National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-55376-6

Canada

THE UNIVERSITY OF ALBERTA

TOWARDS THE AUTOMATION OF CLASSIFICATION PROCEDURES IN CHOROPLETH MAPPING

BY

CLAIRE LAROSE

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF GEOGRAPHY

EDMONTON, ALBERTA

FALL 1989

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: CLAIRE LAROSE

TITLE OF THESIS: TOWARDS THE AUTOMATION OF CLASSIFICATION PROCEDURES IN CHOROPLETH MAPPING

DEGREE: MASTER OF SCIENCE

YEAR THIS DEGREE GRANTED: FALL 1989

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

Claire Larose
(Student's signature)

2069 DELORIMIER
(Student's permanent address)
LONGUEUIL, QC
J4K 3P1

Date: October 12, 1989

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Towards the Automation of Classification Procedures in Choropleth Mapping, submitted by Claire Larose, in partial fulfilment of the requirements for the degree of Master of Science.

(Supervisor)

Date: 1989 October 6

Abstract

The automation of choropleth mapping requires that cartographers identify and formalize regularities in map design decisions. One important aspect of design is the classification of data. This research examines data classification procedures to determine if a given type of frequency distribution may be consistently associated with a particular classing method. Eight classing methods were examined: quantiles, equal steps, arithmetic progressions, geometric progressions, reciprocal progressions, nested-means, standard deviations, and Jenks' "optimal" system. A selection of thirty-five variables is used to study the classification methods. Three measure indices are employed to estimate the quality of the classification. These measure indices are the Sum of Differences, the Tabular Accuracy Index, and the Goodness of Variance Fit. The results indicate that some methods usually operate more efficiently with a particular type of distribution. However, pronounced leptokurtic distributions seem to be a problem for most of the classing systems. The Jenks' "optimal" method appears to be the best classing system for all types of distributions, including leptokurtic distributions, but requires the data to be multimodal. It is therefore possible to create a classing routine which involves only minimum intervention from the user.

Acknowledgement

I want to thank my advisor, Dr. V.T. Noronha, for his comments and help. I also want to acknowledge the two other members of my committee, Dr. D.B. Johnson and Dr. A. Peterson.

This thesis would not have been undertaken, if it had not been for Dr. André Roy, who transmitted to me his excitement about automated mapping.

I was very fortunate to have Bonnie Gallinger, Janice Hanright and Len Sielecki, without who I would still have my nose in an english dictionary and grammar book. Without the computing knowledge of Dan Hemenway, I would still be in front of a microcomputer fixing routines.

I particularily want to thank all my friends, in Edmonton, Montréal, Jasper, Victoria, Vancouver, and even in Germany, who always gave me great moral support and so much more.

I am especially grateful to my parents and my brother. Despite the distance, they never failed to give me moral support and love through their letters and numerous phone calls.

Funding for this thesis was provided by the Natural Sciences and Engineering Research Council of Canada and the University of Alberta (Graduate Faculty Fellowship).

Table of Contents

Abstract iv

Acknowledgement v

Table of Contents vi

List of Tables viii

List of Figures ix

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: DEFINITION OF CHOROPLETH MAPPING 3

2.1 DEFINITION OF MAPPING AND CHOROPLETH MAPPING 3

2.2 CHOROPLETH MAPPING PROCEDURES 4

2.2.1 Preliminary Decisions 4

2.2.2 Classification Procedures 5

2.2.3 Polygon Fill Procedures 6

2.2.4 Finalizing Procedures 7

CHAPTER 3: CLASSIFICATION PROCEDURES 11

3.1 INTRODUCTION 11

3.1.1 Class Rules 14

3.1.2 Number of Classes 15

3.2 CLASSIFICATION METHODS AND CLASS EVALUATION 16

3.2.1 Types of Data Distributions 16

3.2.2 Classification Methods 19

3.2.2.1 Quantiles 20

3.2.2.2 Equal Steps 21

3.2.2.3 Arithmetic, Geometric, and Reciprocal Progressions . . 21

3.2.2.4 Nested-Means and Standard Deviations 23

3.2.2.5 "Optimal" 23

3.2.3 Evaluation Measures 24

3.3 CLASSIFICATION PROCEDURES AND THE COMPUTER 26

CHAPTER 4: EXPERIMENT PLAN 28
4.1 OBJECTIVE OF THE RESEARCH 28
4.2 DATA ACQUISITION 28
4.3 RESEARCH METHODOLOGY 32

CHAPTER 5: EVALUATION OF CLASSIFICATION METHODS 34
5.1 QUANTILES 35
5.2 EQUAL STEPS 38
5.3 ARITHMETIC PROGRESSIONS 41
5.4 GEOMETRIC PROGRESSIONS 44
5.5 RECIPROCAL PROGRESSIONS 47
5.6 NESTED-MEANS 50
5.7 STANDARD DEVIATIONS 50
5.8 "OPTIMAL" 53
5.9 CONCLUDING COMMENTS 58

CHAPTER 6: CONCLUSION 60
6.1 TOWARDS THE AUTOMATION OF CHOROPLETH MAPPING? ... 61

Bibliography 63

Appendix A 66

List of Tables

Table 4.1 List of variables 30

Table 4.2 Skewness and kurtosis for each variable 31

List of Figures

Figure 2.1 Arrangements of dots 7

Figure 2.2 Schema of choropleth mapping procedures 9

Figure 3.1 Unclassed choropleth map 12

Figure 3.2 Types of distributions 17

Figure 5.1 Female population between 55 and 64 years old (quantiles method) 36

Figure 5.2 Choropleth map with classification defined by the quantiles system 37

Figure 5.3 Female population between 55 and 64 years old (equal steps method) 39

Figure 5.4 Choropleth map with classification defined by the equal steps system 40

Figure 5.5 Female population between 25 and 34 years old (arithmetic progressions method) 42

Figure 5.6 Choropleth map with classification defined by the arithmetic progressions system 43

Figure 5.7 Female population between 25 and 34 years old (geometric progressions method) 45

Figure 5.8 Choropleth map with classification defined by the geometric progressions system 46

Figure 5.9 Female population between 25 and 34 years old (reciprocal progressions method) 48

Figure 5.10 Choropleth map with classification defined by the reciprocal progressions system 49

Figure 5.11 Male population between 35 and 54 years old (nested-means method) 51

Figure 5.12 Choropleth map with classification defined by the nested-means system 52

Figure 5.13 Male population between 35 and 54 years old (standard deviations method) 54

Figure 5.14 Choropleth map with classification defined by the standard deviations system 55

Figure 5.15 Female population 65 years old and over (Jenks' "optimal" method) 56

Figure 5.16 Choropleth map with classification defined by the "optimal" system 57

Appendix A Rank-size graph, frequency histogram and table of results for each variable 66

CHAPTER 1: INTRODUCTION

The growing demand for choropleth maps by private and public organizations has created a definite need for cartographically sound automated mapping. Although choropleth mapping software already exists, the maps produced are not based upon good cartographic principles (Lai, 1986; Noronha, 1987; Turner, 1987). Choropleth maps should be objective, be produced rapidly, offer flexibility (so that modifications can be made with ease) and reproducibility, and be academically valid. It is the responsibility of cartographers to improve the quality of maps produced by automated procedures, by influencing the specifications of the hardware and software that produce these maps.

The first step in the process of producing better choropleth mapping software is to build a model of choropleth mapping. This may be attained by defining objective procedures to produce a choropleth map rather than depending on subjective decisions. By defining the objective procedures for the process of choropleth mapping, a model may be developed and may ultimately be integrated into an automated routine, where user intervention may be minimal.

Choropleth maps are a planimetric representation of statistical distributions with an areal symbology. Procedures to produce a choropleth map can be divided into four major steps: preliminary decisions, classification procedures, polygon fill procedures, and finalizing procedures. The first set of procedures accumulates information for the desired map. This information has a major influence on the map, as it is needed in the following steps. The second set of procedures, classification, groups values to obtain classes. Class limits must divide the frequency distribution in an appropriate way. The third set of procedures areally fills the polygon and a logical progression in the filling is necessary. The last set of procedures considers the details or elements involved to complete a choropleth map. Furthermore, some balance in the arrangement of the design elements and the principal figure must be seen.

Based upon the premise that a choropleth mapping model can be developed and integrated in an automated routine, this research concentrates on classification procedures

in choropleth mapping. It attempts to identify and examine the objective elements of the classification process. There are three major types of decision: whether or not to class, to determine the number of classes, and to choose the most appropriate class limits. The first type of decision, whether or not to class data, is still a source of disagreement among cartographers. The second, the choice of the number of classes, has yet to be objectively done. While some parameters may be used to find the maximum or minimum number of classes, the cartographer still makes the final decision on the number of classes for representing a data distribution. However, for the third decision type, it seems possible to reduce the intervention of the cartographer in the delimitation of class limits. This research attempts to demonstrate that it is possible to do this. Eight classing methods are compared for two different numbers of classes. This comparison may be qualitative (using rank/size graphs, maps) or quantitative (using indices to measure the quality of class limits).

It is important to consider the environment in which this research stands. This research observes every operation in the context of an environment where budget is a concern. Therefore microcomputers are used, limited memory and processing speeds are considered, and at best the output has a medium quality. Furthermore, because of the subject of the thesis in itself, classification procedures, there is an emphasis on information content of the map rather than output quality.

This thesis begins with a definition of choropleth mapping and a description of the procedures required to build choropleth maps. This is followed by a more detailed description of the classification procedures. That section also includes a comment on the classing methods and evaluation measures employed in this research, and briefly reviews what would be involved in setting these procedures into an automated routine. The next chapter gives the objectives of this research and describes the methodology followed. Then, each method is tested and evaluated with different measure indices and visual verification. Finally, concluding comments and observations are presented.

CHAPTER 2: DEFINITION OF CHOROPLETH MAPPING

2.1 DEFINITION OF MAPPING AND CHOROPLETH MAPPING

Maps are the representation of concrete or abstract phenomena, localized in space. A map is a reduced, generalized representation of the earth, on a plane showing the situation, the distribution, and the relationships of natural and social phenomena. Cartography is "the art, science and technology of making maps" (Robinson, Sale & Morrison, 1978, p.3). Mapping can be either general or thematic. General mapping (descriptive cartography) attempts to show the distribution of spatial information, while thematic mapping (analytical cartography) is used as a tool for research (Kretschmer, 1978). Choropleth mapping is one kind of thematic mapping and it produces maps which essentially have an analytical purpose. The emphasis is more on the attribute than on location. Choropleth mapping tends to emphasize spatial relationships, the variations of a variable in space. Therefore, features such as accuracy and functionality are required.

Choropleth is a name which derives from Greek 'choros,' for place, area or space, and 'plethos,' for magnitude, multitude or number. Choropleth maps symbolize quantitative data for areas. They are included in the kind of mapping "which involves some indication of quantity as well as of spatial distribution, known as quantitative areal mapping" (Monkhouse & Wilkinson, 1971, p.39). Choropleth mapping has boundary lines as a characteristic. The locations of the statistics represented coincide with the boundaries of the unit areas (Monkhouse & Wilkinson, 1971; Robinson, Sale & Morrison, 1978). However, no quantity is assigned to these boundary lines. Therefore choropleth mapping could be defined as the kind of mapping which symbolizes, in some manner, a quantitative magnitude (or z-value) that applies to a particular unit of area, usually an administrative region.

Some restrictions have to be considered with choropleth mapping, particularly in the type of data to use. There is a high correlation between area and the number of items found within the areal units. Therefore, choropleth maps created with raw values tend

more to reflect the areal size of a data unit, and confusion between unit size and number of items enumerated is created. To remove the effect of variations in the size of the unit, values must be transformed into density measures, ratios, averages or percentages. Values cannot be absolute or raw (Cuff & Mattson, 1982; Robinson, Sale & Morrison, 1978; Truran, 1975). The only exception is when all areal units have the same size (in the case of population representation) or the same population (for other demographic variables based on population).

2.2 CHOROPLETH MAPPING PROCEDURES

Different procedures are required to construct a choropleth map. Four groups of procedures can be distinguished:

- preliminary decisions
- classification procedures
- polygon fill procedures
- finalizing procedures.

2.2.1 Preliminary Decisions

This procedure cumulates preliminary information about the desired map. This information covers the phenomena that exercise a predominant influence on the map. These phenomena are called "motivation factors" (DeBrommer, 1969). The motivation factors allow the cartographer to determine the content of the map and the choice of the expression modes. These factors are essentially the purpose or the use of the map, the type of audience, and the levels of information quality and readability of the map.

Once it is decided to make a choropleth map, some motivation factors are already answered. For example, it is already known that the map will be an information resource and therefore requires features such as accuracy and functionality. The audience often

tends to be of a scientific nature (analyst or specialist) and the size of this type of audience tends to be smaller than the size of a general public audience. The purpose of a choropleth map can be either of two kinds: analysis (including pattern recognition) or illustration. The first approach implies a direct use of the map. With the second approach the map is only a means of illustration and explanation of an author's research. However, the "illustration" map users tend to be specialists themselves, familiar with the research subject or field. Therefore, in terms of output controls, such a choropleth map may be moderately to highly clear, simple, and the information quality may be of low to moderate importance. If the choropleth map is built with an analytical purpose, the importance of the two output controls is reversed. The design of the map emphasizes information at the expense of the simplicity. To obtain a generalized map clearly showing patterns, the cartographer may prefer a compromise between information and readability, i.e. the two are represented with an intermediate emphasis.

2.2.2 Classification Procedures

Classification procedures group the values to obtain classes. These procedures can be omitted in the map building process if the cartographer intends to draw a choropleth map without class intervals. In this case, each areal unit is assigned its real value rather than a classed value.

To obtain a good data classification, the cartographer must decide on the number of classes and the classification method. The number of classes is essentially a function of the intended use, purpose, and type of audience of the choropleth map. A map used as an analytical tool and by only a small group of researchers (or a single individual) may have a greater number of classes than a map destined for a bigger audience. However, as a general rule, the number of classes should not exceed ten (Burrough, 1986; Dickinson, 1963; Jenks, 1971). To classify a data distribution, the cartographer may agglomerate or subdivide data (Johnston, 1968; Morrison, 1975; Robinson, Sale & Morrison, 1978). The object of agglomeration is to find groups or clusters of similar objects (or spatial units) or

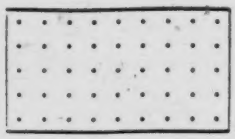
items (Semple & Green, 1984). However, since agglomeration techniques may require more effort and time (computing time and memory space), subdivision methods may be favoured. These methods may use mathematical functions or statistical measures, or operate more arbitrarily or systematically. Generally, the selection of a classing method is a function of the characteristics of the data distribution. Depending on the kind of data distribution, one method may be more appropriate than the others.

Once a method is selected (when subdividing), classes can be generated and some measure indices may be used to check the classification and perhaps move the class limits slightly to get a better approximation to natural class breaks, if any (natural breaks will be defined in Chapter 3).

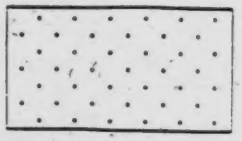
2.2.3 Polygon Fill Procedures

The polygon fill procedures consider in which manner the areal units of the choropleth map will be represent. For the polygon fill, it is important to use a logical progression of shading or colours. Lower values should be represented by a lighter fill and higher values should be represented by a darker fill. Three kinds of polygon fill are available: patterns, black and white (grey scale), and solid colour. Dots and lines are the two essential features of an apparent structure fill. Lines may be parallel or crossing, with or without an angle. There are three different arrangements of dots: square, oblique square, and triangular (Figure 2.1). Only the triangular arrangement offers an even distance in all directions between the dots. The quantitative progression from one category to another is expressed by a variation in line thickness or dot size, the spacing between the lines or dots, or a combination of the two, i.e. a change in size and spacing.

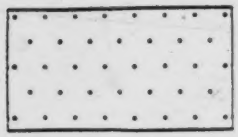
Figure 2.1 Arrangements of dots



a) Square



b) Oblique square



c) Triangular

To fill a polygon with colour or in black and white, it is important to know the three characteristics of colour: hue, chroma, and value. Hue is the obvious feature of monochromatic light that varies with wavelength. It is used to denote various regions of the spectrum. Chroma, or saturation, is the degree of brightness or purity of a hue. Chroma refers to the lack of "whiteness" in a colour, or how much a colour differs from white (or grey). Ultimately, all colours, with a chroma of 0, become grey. In general, on a choropleth map, chroma tends to be the same for all categories. Value is the lightness or darkness of a colour as rated on a grey scale from white to black. It is the attribute that describes the perceived intensity of light. Value is the characteristic that changes for each category. In black and white, the polygon fill is characterized by a series of grey tones which give visually equal appearing steps from white to black. This is generally called a grey scale. In colour, the principle remains the same.

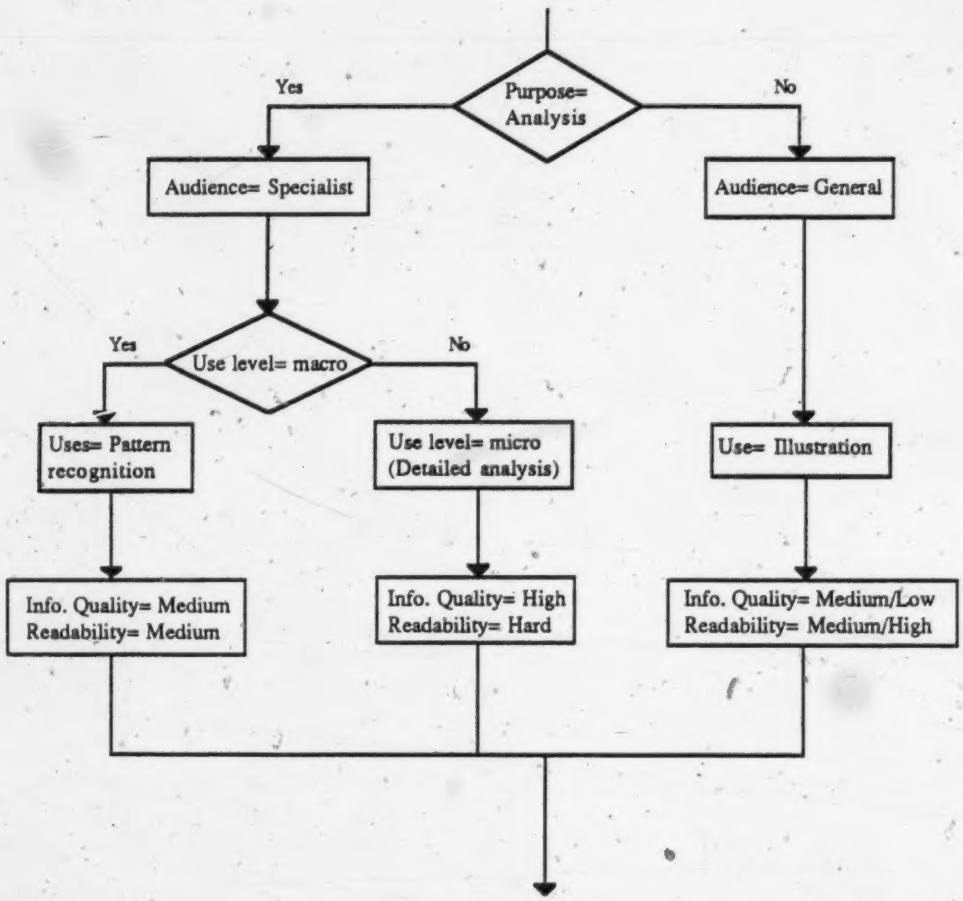
2.2.4 Finalizing Procedures

These procedures are concerned with the details or elements involved to complete a choropleth map. Some balance in the arrangement of the principal figure and of the design elements needs to be performed. The design elements are essentially the scale, the title, the legend, the data sources and, if necessary, the orientation. To balance all the

elements of the map, the size, the orientation, and the location of the principal figure and of the design elements must be considered. The importance or the necessity of every element on the map must also be examined, since the map must always remain as simple and clear as possible. The legend is an important component of the map; it should be clear and easy to understand; if possible, round numbers should be used, and all classes represented on the map should be defined.

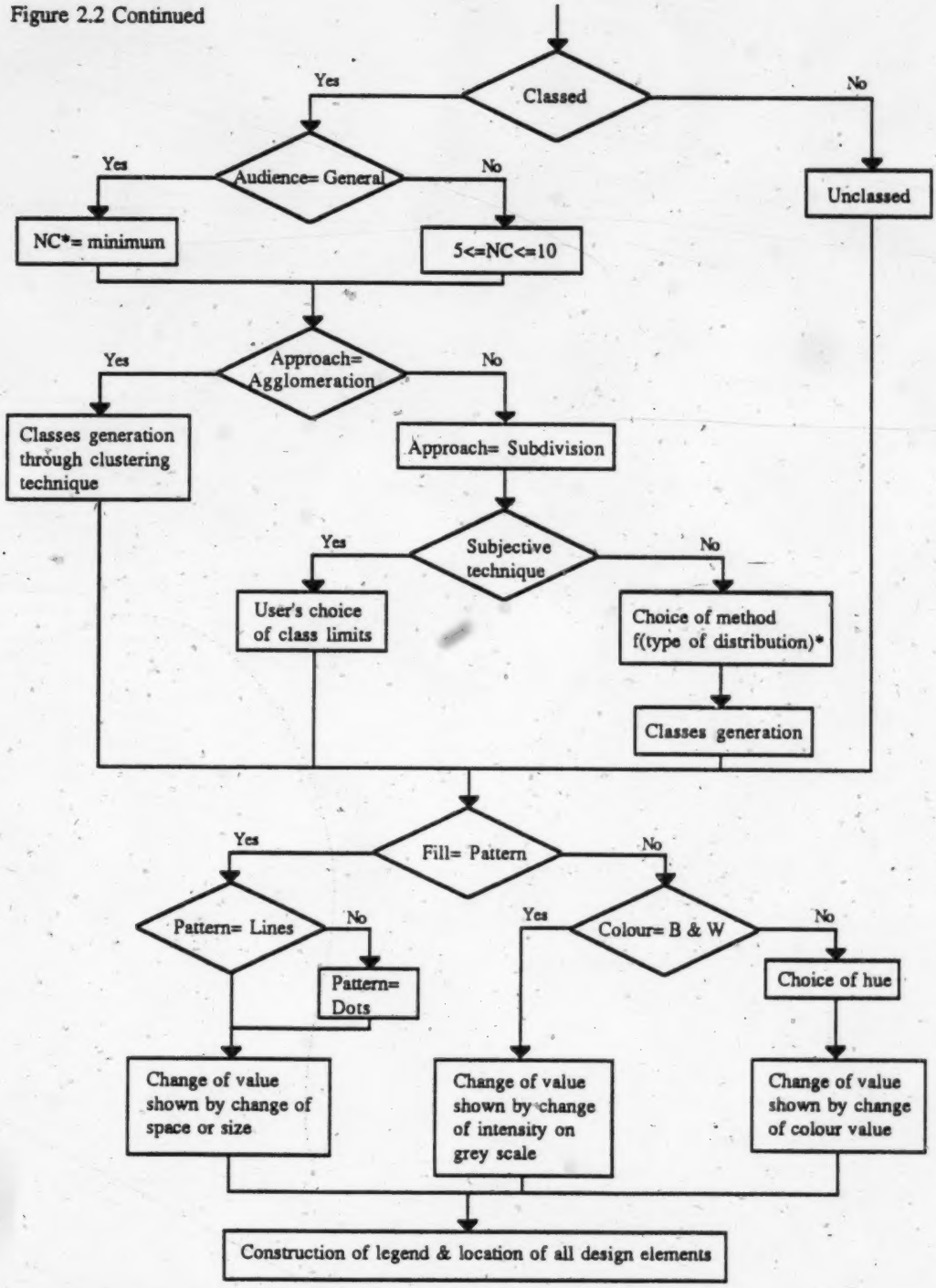
The four procedures described above, are all effectuated in a certain order. The sequence goes as follows: the cartographer determines under which circumstances the choropleth map will be employed; once the motivation factors are defined and the data acquired, classification of the data is effectuated and classes are generated; for each category, a certain type of polygon fill is assigned, following a logical progression; finally, the design elements are added and the cartographer ensures that all the elements are relatively well balanced on the choropleth map (Figure 2.2).

Figure 2.2 Schema of choropleth mapping procedures



Continued on next page

Figure 2.2 Continued



* NC: Number of classes
f(): Function of ()

CHAPTER 3: CLASSIFICATION PROCEDURES

3.1 INTRODUCTION

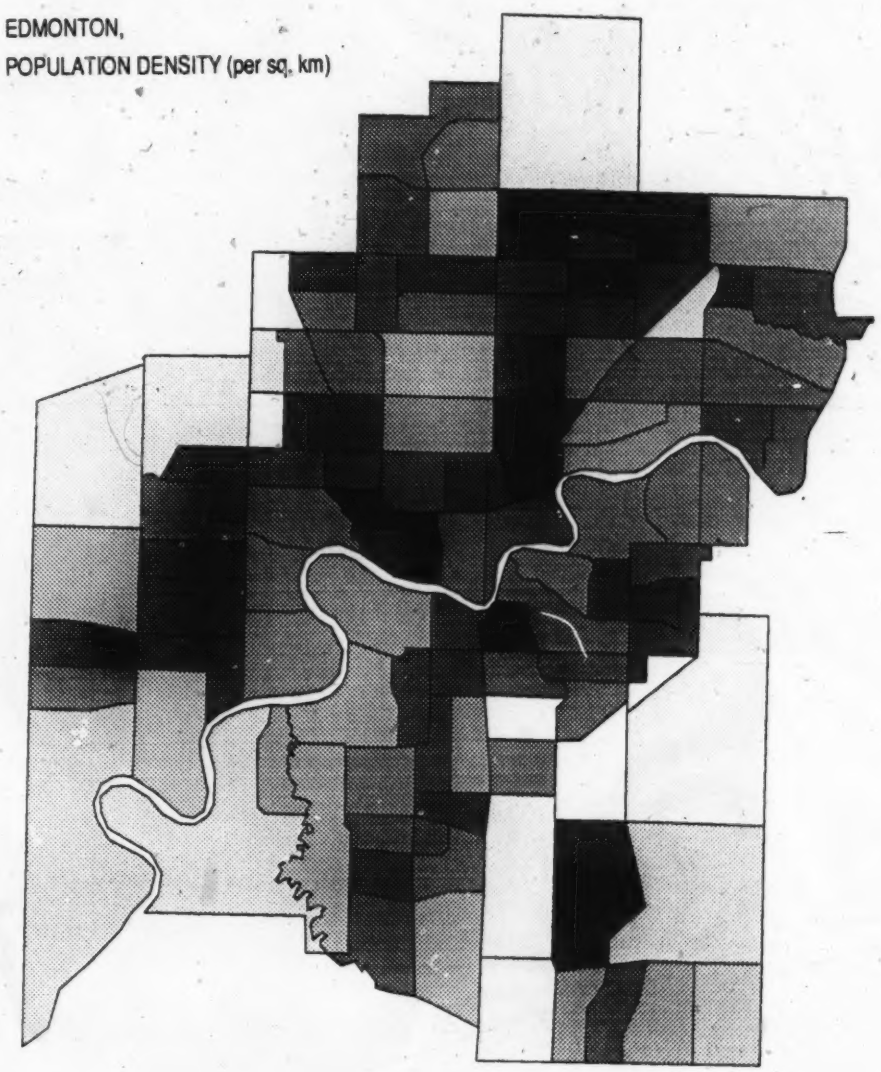
Classification procedures arrange data into groups to produce some kind of categorization. This grouping is based on some similarity of properties or on relationships between objects (Semple and Green, 1984). Since data are grouped to simplify a map, classification is often considered as a form of generalization. This generalization of data becomes important when the principal objective of choropleth mapping is to show patterns of distribution. Therefore, classification may be defined as the "standard intellectual process of generalization that seeks to sort phenomena into classes in order to bring relative order and simplicity out of the complexity of incomprehensible differences, inconsequential differences, or the unmanageable magnitudes of information" (Robinson, Sale & Morrison, 1978, p.152).

Fisher (1982) states the following two main reasons why data should be classified:

- to reduce the volume of detail work required of the mapmaker;
- to improve overall comprehension through the use of fewer numbers and distinctions.

Classification is often necessary to reduce the complexity of a map and to increase the user's ability to understand and assimilate the information. In addition, it can be used to extract the general patterns of a data distribution (Caldwell, 1981; Cuenin, 1972; Jenks, 1970).

In 1973, Tobler proposed a choropleth map without classed values. The choropleth maps showed visual intensity exactly proportional to data intensity. Initially, the symbology on unclassed choropleth maps consisted of crossing lines. The spacing between lines changed proportionally to data intensity. With current technology, it is possible to represent intensity values with different tones on a grey scale (e.g. Figure 3.1). The principal reason for producing unclassed choropleth maps is that, since there are



Source: Statistics Canada 1981

Figure 3.1 Unclassed choropleth map

no class intervals, there is no quantization error. Accuracy is one of the major advantages of no-class mapping.

However, this approach is controversial. Opponents of unclassed mapping argue that the lack of classification inhibits readers' ability to recognize information. As more information is represented on a map, its clarity decreases and it becomes more complex to interpret. As a result, more effort from the map reader is required to assimilate the information (Brassel, 1979; Caldwell, 1981; Muller, 1979). Opponents also argue that unclassed mapping may not be a regionalization of a variable, and therefore may not meet one principal goal of choropleth mapping, which is to show patterns of distribution. Referring to an important characteristic of choropleth maps, generalization, Stefanovic and Vries-Baayens (1984, p.53) state that "accuracy cannot be the primary design objective."

Proponents of unclassed mapping indicate that accuracy is definitely increased (Muller, 1979; Tobler, 1973), and that map readers can see and reorganize the elements of an unclassed map in a consistent and logical fashion, to extract patterns of the distribution (Cuff, 1982; Muller, 1979). Map readers can make more accurate estimations of mapped values since the process of grouping data can distort the information (Cuff, 1982). Another advantage of unclassed choropleth mapping, according to its proponents, occurs at the edges of each areal unit. Choropleth maps assume homogeneity within each areal unit and often the change from one unit to another is abrupt. With unclassed maps, this transition may be less abrupt, and estimation of the distribution within an areal unit may be easier, by comparing it with its neighbours. However, it may also be an error to assume that this transition is not abrupt. A geographical feature, acting as a barrier, may coincide with the limits of two neighbouring units. In such a case, characteristics of both populations may be different and the change may be abrupt in reality. A final argument for unclassed mapping, by Muller (1979), is that visual discrimination should be higher in importance than visual identification. A choropleth map should be seen and not read, therefore the identification of the class to which each unit belongs is irrelevant. However, Tobler proposed no-class mapping with the idea of inversion: a map user should be able

to find the exact numerical value for any areal unit simply by extracting it from the map. Under such conditions, the purpose of the map is to be both read and seen. Following the inversion concept of Tobler (1973), a numerical table seems more appropriate and certainly involves less work for the user. The choropleth map is not an illustrated table, but a means to show patterns of distribution.

Whether or not unclassed choropleth maps should be used is still a concern and it does not seem that a solution to this problem has been found. While some cartographers are trying to find techniques to define optimal class limits for a data distribution, others believe that unclassed maps produce results as good as, if not better than, classed maps.

In the context of this research, in order to schematize choropleth mapping procedures for a general audience, classed choropleth mapping may appear to be the appropriate choice. For a more specialized audience, the choice between classed and unclassed mapping may remain in the hands of the cartographer or the specifications of the user. For the purpose of pattern recognition, a choropleth map with classes may be more suitable, while for detailed analysis or to locate data, unclassed mapping may be a good solution. Since choropleth mapping is a generalization of data which can vary in degree, the map user should not face a map that is too complex.

3.1.1 Class Rules

Choropleth maps should give a good representation of the characteristics of the data distribution and show good patterns of distribution. To be cartographically sound, choropleth maps should follow some criteria or rules in the classification process (Coulson, 1987; Davis, 1974; Fisher, 1982; Jenks & Coulson, 1963):

- the class intervals should cover the full range of the data;
- no value should occur in more than one class, i.e., the classes should not intersect;
- no vacant class should appear;
- there should be minimum variation within classes while there should be

maximum variation between classes.

Therefore, these criteria may be used to judge the quality of the class limits of a data distribution.

3.1.2 Number of Classes

Studies on perception in mapping have demonstrated that map readers usually cannot distinguish more than nine or ten different patterns on a map (Dickinson, 1963; Jenks, 1971). In general, cartographers agree that choropleth maps should not have more than eight to ten classes. Meanwhile, a map with only two classes may not provide enough information to make worthwhile analysis. Therefore, it is possible to select between five and ten classes as an appropriate number of categories for a choropleth map (Burrough, 1986).

Although this procedure is of great importance, there is no established approach for determining the most appropriate number of classes. Clustering techniques may be one solution. A cluster of points identifies a subset of items which are more similar to each other than to other items outside the cluster. The similarity of any items with respect to the variables involved is measured by the distance between these points. Following the clustering principle, Monmonier (1973) employs a method using eigenvalues to find natural clusters in a data distribution in order to determine the most appropriate number of classes. However, this technique requires a substantial amount of computer time and memory. Davis (1974) and Monkhouse and Wilkinson (1971) discuss an objective approach for determining a first approximation for the number of classes that should be chosen. It appears that the maximum number of classes (c) should not exceed five times the logarithm of the number of observations (n):

$$c = 5 \cdot \log_{10}(n) \qquad (3.1)$$

At times, a large number of classes may introduce the problem of empty categories. Gaps between values may also increase the risk of having empty classes in a data classification.

In general, the intended use of the map, and implicitly the type of audience, determines the choice of the number of classes. A choropleth map, created for an illustrative purpose and for a relatively general audience, may require a minimum number of categories. This type of map needs to be simple and clear. If the intended use of the map is for analysis (specialist users), the number of classes may be increased. However, as a general rule, this value may vary between 5 and 10. If the analysis is strictly pattern recognition, the number of classes may be lower. Ultimately, the final decision of choosing the number of categories still remains with the cartographer.

3.2 CLASSIFICATION METHODS AND CLASS EVALUATION

3.2.1 Types of Data Distributions

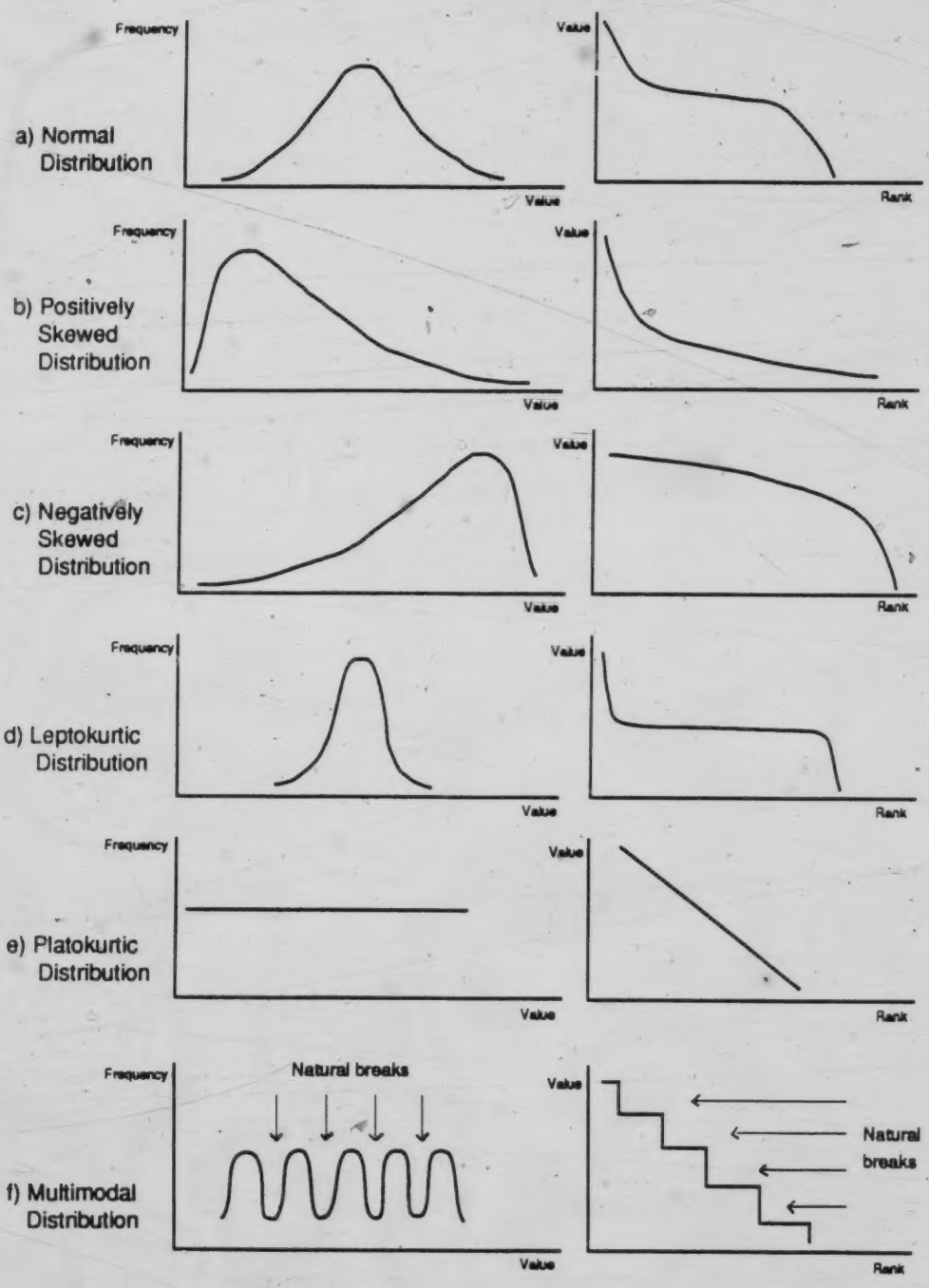
Different types of data distributions exist. It is important to know how to recognize them and to be able to find the most appropriate classing method. Each type of distribution is characterized by a particular shape, and this shape can be described in terms of skewness and kurtosis indices. Skewness "measures the extent to which the bulk of the values in a distribution are concentrated to one side or the other of the mean" (Ebdon, 1985, p.30). Skewness is calculated as follows:

$$\text{Skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n \cdot \sigma^3} \qquad (3.2)$$

where $\bar{x}$ =distribution mean,
n =number of observations,
σ =standard deviation,
x_i=value at i (Ebdon, 1985).

When the distribution is perfectly symmetrical around the mean, the skewness is zero (Figure 3.2.a). A distribution is skewed when the peak or mode of the frequency distribution does not correspond to the mean. The bulk of the distribution is on one side or the other of the mean. If most of the values are less than the mean, the distribution is

Figure 3.2 Types of Distribution



positively skewed (Figure 3.2.b). The skewness is greater than zero. If there are more values greater than the mean, the distribution is negatively skewed and the skewness is negative (Figure 3.2.c).

Kurtosis measures the extent to which values are concentrated in one part of a frequency distribution — it refers to the notion of peakedness in a frequency distribution. Kurtosis is calculated as follows:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n \cdot \sigma^4} \qquad (3.3)$$

Frequency distributions may also be very peaked or very flat. Generally, a kurtosis of 3 represents a normal distribution. If the kurtosis is above 3 the distribution is leptokurtic; there is a major peak (Figure 3.2.d). A distribution is leptokurtic if one class, or a group of adjacent classes, in a frequency distribution contains a large proportion of all the values in the distribution. A platokurtic distribution, characterized by a kurtosis below three, is a flat distribution (Figure 3.2.e). Each class contains approximately the same proportion of all the values. However, flat distributions are almost never encountered since most geographical variables are positively skewed and leptokurtic (Ebdon, 1985; Silk, 1979).

The previous types of distributions are described in a unimodal form (single peak on a frequency distribution histogram). However, often a frequency distribution has more than one peak and is multimodal (Figure 3.2.f). An abrupt change of gradient in a frequency distribution is characteristic of multimodal distributions. The value where this abrupt change occurs is often called a natural break. Since the groups of values on each side of a natural break tend to represent values with most similarities, natural breaks are often used as class limits.

3.2.2 Classification Methods

Evans (1977) mentions sixteen classing systems. These may be categorized into four groups: exogenous, arbitrary, idiographic, and serial. Briefly, their characteristics are as follows.

Exogenous methods define fixed class intervals that "are fixed according to threshold values that are relevant to, but not derived from, the data set under study" (Burrough, 1986, p.137). This type of method can be useful in map series.

Arbitrary methods emphasize simple class limits, which are easily understood and have no reference to the data distribution. The class limits are usually round numbers and often the intervals are regularly spaced.

Idiographic techniques choose class limits with respect to specific characteristics of the data set. These systems usually produce irregular class limits. They include multimodal, multi-step, contiguity-biased, correlation-biased, quantiles, and nested-means methods.

Serial techniques define class intervals by the use of a mathematical function. The class limits are determined in relation to statistics for the overall frequency distribution (e.g. median, mean, standard deviation, and range), but not to individual details of the distribution. Class limits are generally regularly spaced. Normal percentiles, standard deviations, equal arithmetic intervals with no variation in class width, equal intervals on a reciprocal scale or on trigonometric scales, geometric progressions, arithmetic progressions and curvilinear progressions, are all serial methods.

Since neither exogenous nor arbitrary methods consider the distribution of data, idiographic and serial methods may be better alternatives (Evans, 1977). Eight classing systems are compared in this research. They have been selected because of their popularity within the field of cartography or on different authors' recommendations. Three systems are idiographic: quantiles, nested-means, and Jenks' "optimal" method. The other methods are serial: standard deviations, arithmetic progressions, geometric progressions, reciprocal progressions, and equal steps.

There are two additional classing methods, not mentioned by Evans, which consider

geographical area when defining class limits: the clinographic curve and the cumulative frequency curve. Both methods require visual examination of a graph. The Y-axis is scaled arithmetically and represents the values of data. The X-axis shows the cumulative areas. For the clinographic curve, the X-axis is scaled in percent of total area, with a square root scale from 0 to 100% (Robinson, Sale, & Morrison, 1978). The X-axis is scaled arithmetically for the cumulative frequency curve. The critical points are the points where the slope of the curve changes. In both cases, class limits correspond to the critical points. However, these two systems rely heavily on the intervention of the cartographer since critical points are located through a visual inspection of a curve, and this is why they are not examined in this research.

3.2.2.1 Quantiles

The quantiles method divides data into equal frequencies and does not consider individual values. To define class limits the data are first sorted. The total number of areal units is then divided by the number of classes desired, and the result gives the frequency that should appear in each class. When the total frequency cannot be evenly divided by the number of classes, the remainder are distributed among the lower categories (Chang, 1974). The quantiles method may be preferred with data which cannot be easily classified or when the type of frequency distribution is not evident (Davis, 1974). This method ensures that all classes are approximately equal in importance (Davis, 1974; Evans, 1977).

The following notation is used in the discussion below:

n =number of observations,
c =number of classes,
f_i=number of units in a class i, i=1,2,...c,
s =remainder of n÷c
U_i=upper class limit for class i, i=1,2,...c,
L_i=lower class limit for class i, i=1,2,...c,
x_j=value of data item j, j=1,2,..n,
M =maximum of the frequency distribution,
m =minimum of the frequency distribution.

The class limits for the quantiles method are defined as follows:

$$f_i = \begin{cases} INT(\frac{n}{c})+1 & \text{for } i = 1,2,...,s \\ INT(\frac{n}{c}) & \text{for } i = s+1,...,c \end{cases} \quad (3.4)$$

where INT is a function that returns the largest integer less than or equal to the argument.

$$U_i = x_j \qquad \text{where } j = \sum_{q=1}^{i} f_q$$

3.2.2.2 Equal Steps

The equal steps or equal intervals method employs data range. The data are divided into equal classes and class limits are determined with this formula:

$$r = M-m \quad (3.5)$$

$$W = \frac{r}{c}$$

$$U_i = m+(W \cdot i)$$

where r=range of the frequency distribution,
W=class width (constant).

Although class limits are easy to calculate this system does not provide great correspondence between classing accuracy and skewness (Davis, 1974; Smith, 1986). Equal intervals may be more appropriate for flat frequency distributions.

3.2.2.3 Arithmetic, Geometric, and Reciprocal Progressions

For the three following methods, arithmetic, geometric and reciprocal progressions, a mathematical relation between class limits is followed. All three techniques initially calculate the rate of increase of classes frequencies and use it to determine class limits. These methods may be useful for skewed distributions, since they find intervals which become systematically smaller toward either the upper or lower end of the scale. They try "to preserve the regularity of intervals as much as possible and at the same time they try to

avoid a too heavy clustering of data in individual classes" (Stefanovic & Vries-Baayens, 1984, p.55). Truran (1975) states that arithmetic progressions can be useful when the range of values is not great. Arithmetic progressions determine class limits as follows:

$$Z = \frac{r}{(1+2+...+c)} \tag{3.6}$$

Z is interpreted as the rate of increase of class frequencies,

$L_1 = m$

$L_i = L_{i-1}+(i \cdot Z)$ for i=2,3,...c.

Geometric progressions may be more useful because they allow a wider range of values to be represented and at the same time give greater consideration to the lower values (Truran, 1975).

$$Z = \frac{\log(M)-\log(m)}{c} \tag{3.7}$$

$U_i = 10^{\log(m)+(i \cdot Z)}$ for i=1,2,...(c-1)

$U_c = M$

For very positively skewed data distributions, reciprocal progressions may be a good alternative. This technique defines narrow class limits at the lower end of the range. However, Davis (1974) recommends that reciprocal progressions be used only with continuous data.

$$Z = \frac{\frac{1}{m} - \frac{1}{M}}{c} \tag{3.8}$$

$$U_i = \frac{1}{\frac{1}{m}-(i \cdot Z)} \quad \text{for i=1,2,...(c-1)}$$

$U_c = M$

3.2.2.4 Nested-Means and Standard Deviations

Nested-means and standard deviations methods should be employed with normally distributed sets (Paslawski, 1984; Robinson, Sale & Morrison, 1978; Chang, 1974). Both approaches are based on statistical properties of the data (i.e. the mean or the standard deviation). With the nested-means system, the mean of the distribution is used as a point of division to give two classes. Each of these classes may be subdivided at its own mean, and the process may be pursued with these new means. Therefore, the number of classes always equals 2^k, where k is the number of levels of subdivision.

Standard deviations on the other hand define class width as fractional or integral multiples of the standard deviation. Class intervals are above and below the mean.

Smith (1986) and Stefanovic & Vries-Baayens (1984) consider these two methods unreliable for any type of distribution: the class intervals they produce can be regular or irregular, but this is unpredictable, and the relationship with skewness and kurtosis is not always strong.

3.2.2.5 "Optimal"

The "optimal" method (Jenks, 1977) selects class limits on the basis of variance or of absolute deviations, after a certain number of iterations. A first set of class intervals is defined and an objective function — either the pooled within-group sum of squares (PWGSS) or absolute deviations from the class mean (DCM) — is calculated:

$$DCM = \sum_{j=1}^{c} \sum_{\substack{i=1 \\ i \in j}}^{n} | x_i - \bar{X}_j | \qquad (3.9)$$

where c=number of classes,
n=number of observations,
$\bar{X}_j$=mean of class j.

$$PWGSS = \sum_{j=1}^{c} \sum_{\substack{i=1 \\ i \in j}}^{n} (x_i - \bar{X}_j)^2 \qquad (3.10)$$

An observation at the edge of one class is moved to an adjacent class and the new PWGSS or DCM is calculated. If this transfer reduces the objective function, the observation is left

in its new class. Otherwise, the observation is moved back to its original class. This process is repeated until the objective function is minimized. Since the "optimal" method by DCM gives similar results to the PWGSS method, comments are made only for the latter. Smith (1986), suggests that this technique gives the best results. A weakness of the method, as originally developed by Jenks, is that class limits are not necessarily contiguous. This contravenes cartographic conventions and can become a source of confusion when the map user refers to the legend or simply tries to understand the map. However, this can be corrected very easily by finding an intermediate value between given limits.

3.2.3 Evaluation Measures

To analyze results of the different classing techniques described above, three measure indices are employed: Sum of Differences (Jenks & Coulson, 1963), Tabular Accuracy Index (Jenks & Caspall, 1971) and Goodness of Variance Fit (Smith, 1986).

The Sum of Differences index compares the actual weighted range of each class to a theoretical one (median).

$$\text{Sum of Differences} = \sum_{i=1}^{c} \left(\frac{R_i}{\bar{X}_i} - \frac{R_i}{\frac{R_i}{2} + m_i} \right) \qquad (3.11)$$

where c=number of classes,
R_i=range of class i,
$\bar{X}_i$=mean of class i,
m_i=minimum of class i.

It is assumed that class limits should correspond to natural class breaks, and that the data distribution for each class adopts approximately a normal shape. The Sum of Differences index has a major weakness. Since the calculation of this index implies divisions, when the denominator equals zero (e.g. when a class mean equals zero), the equation becomes meaningless.

The Tabular Accuracy Index (TAI) is derived from another index, the Tabular Error. The Tabular Error is the "discrepancy between the category mean represented cartographically and the value for an areal unit to be found by consulting a numerical, or

tabular, representation of the data" (Monmonier, 1982, p.17). To calculate the Tabular Accuracy index, the Tabular Error index is substracted from 1.0 (the maximum result), and the values of the deviations from the class mean (DCM) and of the deviations from the array mean (DAM) are found.

$$TAI = 1 - \frac{DCM}{DAM} \qquad (3.12)$$

where $DAM = \sum_{i=1}^{n} |x_i - \bar{x}|$

n=number of observations,
$\bar{x}$=array mean.

$$DCM = \sum_{j=1}^{c} \sum_{\substack{i=1 \\ i \in j}}^{n} |x_i - \bar{X}_j|$$

c=number of classes,
$\bar{X}_j$=mean of class j.

The Goodness of Variance Fit (GVF) uses squared deviations rather than simple absolute deviations. Coulson (1987) refers to it as the best test to compare sets of classes and to determine which set divides a distribution most adequately. The GVF index can use either a base of 1 or a base of 100. The GVF index divides the "Sum of Squares Within" by the "Sum of Squares Total" to measure the degree of classing accuracy. The "Sum of Squares Within" (or squared deviations from the class mean, SDCM) sums the squared deviations of class data values from the mean of each class, and then it sums these values over all classes. The "Sum of Squares Total" (or squared deviations from the array mean, SDAM) sums the squared deviations of individual data values from the overall mean of the data set. Therefore, the formula for the Goodness of Variance Fit index is as follows:

$$GVF = \left(1 - \frac{SDCM}{SDAM}\right) \cdot 100 \qquad (3.13)$$

where $SDAM = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$SDCM = \sum_{j=1}^{c} \sum_{\substack{i=1 \\ i \in j}}^{n} (x_i - \bar{X}_j)^2$$

Both TAI and GVF indices operate in a similar way. They both also have a weakness. Unless there is a verification procedure, they do not detect empty classes and still may produce seemingly reasonable results.

3.3 CLASSIFICATION PROCEDURES AND THE COMPUTER

This research looks at the relationship between frequency distribution types and classing methods. It is assumed that such a relation exists, and that ideally a computer routine, which objectively selects the appropriate classing method, can be generated. Such a computer routine would require only a minimum intervention of a user and would produce cartographically sound classifications.

Once the decision is made whether or not to classify, the classification process is straightforward. While most of the procedures may be automated, the choice of the number of classes still needs the intervention of the cartographer. A clustering technique may be employed to determine the number of classes. However, the intended use of the map and the type of audience are ignored in the process, and these are two significant variables to consider when choosing the number of classes.

Since some classing methods may operate more efficiently with particular types of distributions, the characteristics of the frequency distribution must be known. Skewness and kurtosis can give this kind of information and a computer routine must calculate their value. With these two indices, it is possible to know approximately what shape the frequency distribution has: normal, positively or negatively skewed, peaked or flat. To verify if a frequency distribution is reasonably normal, the Kolmogorov-Smirnov test can be used (Barber, 1988). When the shape of the frequency distribution is known, a classing method can be selected to define class limits. Different classing methods would be available in this computer routine. Cartographic rules must be followed and it must be ensured that there are no empty or overlapping classes, and that all data are covered by the classification.

To verify if the class limits are a good approximation of the natural breaks in a frequency distribution, measure indices are calculated. The computer routine can easily include this calculation. When the measure index has a low value, the computer routine (or the cartographer) may either modify the class limits or use another classing method.

In the scenario above, the assumption is made that the essential decisions on data classification and design are formalized to the point that they can be automated. The cartographer is still supervising the process and can override the decision of the computer. The cartographer may be guided by computer-generated statistics such as the measure indices described earlier, but is still the one making the final decision.

The next chapter documents a series of experiments, to study the outcomes of applying the classification procedures and measure indices described above, to realistic data.

CHAPTER 4: EXPERIMENT PLAN

4.1 OBJECTIVE OF THE RESEARCH

The objective of this thesis is ultimately to identify regularities in choropleth map design decisions. These regularities would lead to the creation of a model which could be incorporated in an automated system producing cartographically sound maps. In order to generate such a choropleth mapping model and programme, it is necessary to examine thoroughly the four major mapping procedures: preliminary decisions, classification procedures, polygon fill procedures, and finalizing procedures. This thesis concentrates on the classification procedures. An attempt is made to find a relationship between classing methods and types of frequency distributions. Therefore, this research attempts to determine if one classing method can be more appropriate than another for a particular type of frequency distribution. If this idea tested true, it would mean that the selection of a classing method could be done objectively, and therefore automatically. In the context of a classification routine, less subjectivity (less intervention from the user) is preferred to avoid cartographic errors on the choropleth map. This research also determines if one method can be better than any other for any type of distribution. Therefore, tests are conducted with eight classing methods, relatively efficient or popular among cartographers, on different types of frequency distributions.

4.2 DATA ACQUISITION

The City of Edmonton is used as an example of how the classification procedures can be applied. Edmonton is chosen because of the author's familiarity with the city. The 125 census tracts in Edmonton were digitized from a 1981 UTM base map. Different types of demographic information were gathered from Statistics Canada (1981) for each of Edmonton's census tracts. The variables chosen cover a variety of distributions, and the

problems involved in the construction of choropleth maps. Table 4.1 gives a list of the thirty-five variables present in the data base and they are represented graphically in the Appendix A. Most of the variables are expressed as a percentage of the total population of each areal unit, since choropleth mapping requires that the values be spatially intensive.

$$X_i = \frac{x_i}{p_i} \cdot 100 \qquad (4.1)$$

where x_i=value at i, p_i=population of observation i. Three variables are not expressed as a percentage: "Density," "Average Income Male" and "Average Income Female."

Table 4.2 shows the skewness and kurtosis for each variable. According to the skewness index, nine variables, with a skewness lower than 0.20, can be characterized as "normal" or at least close to normal. They are "Density," "Catholic," "Born in Canada," "Male 0-14," "Male 15-19," "Male 35-54," "Female 0-14," "Female 35-54," and "Female 55-64." These variables have varied kurtoses (leptokurtic, platokurtic or normal). Only the variables "Male 15-19" and "Female 35-54" have normal skewness and normal kurtosis. When skewed, most of the variables are positively skewed. With a skewness index less than -2, only one variable, "Female, Total," could be considered as definitely negatively skewed. One variable, "Grades 9-13," is characterized by a relatively flat frequency distribution. Therefore the variables sampled exhibit these different types of shape: normal or relatively normal, positively and negatively skewed, relatively peaked or flat.

Table 4.1 List of variables

1	Population density per square kilometre	
2	Religion	Catholic
3	Religion	Protestant
4	Religion	Eastern Orthodox
5	Place of Birth	Born in Canada
6	Place of Birth	Born in Alberta
7	Place of Birth	Born outside Canada
8	Place of Birth	Born in Canada, outside Alberta *
9	Scholarity	Less than Grade 9
10	Scholarity	Grades 9-13 (with and without secondary certificate) *
11	Scholarity	University (with and without degree) *
12	Average Income	Male
13	Average Income	Female
14	Male Population	Total
15	Male Population	0-14 years *
16	Male Population	15-19 years
17	Male Population	20-24 years
18	Male Population	25-34 years
19	Male Population	35-54 years *
20	Male Population	55-64 years
21	Male Population	65 years and over *
22	Female Population	Total
23	Female Population	0-14 years *
24	Female Population	15-19 years
25	Female Population	20-24 years
26	Female Population	25-34 years
27	Female Population	35-54 years *
28	Female Population	55-64 years
29	Female Population	65 years and over *
30	Marital Status	Single (15 years and over)
31	Marital Status	Married
32	Mother Tongue	English
33	Mother Tongue	French
34	Mother Tongue	Ukrainian
35	Mother Tongue	German

* Variable created by combination of other variables

Table 4.2: Skewness and Kurtosis for each variable

Variable	Skewness	Kurtosis
(1) Density	.16	3.76
(2) Catholic	-.00	3.52
(3) Protestant	-.63	3.03
(4) Orthodox	1.07	3.97
(5) Born in Canada	-.16	5.65
(6) Born in Alberta	-.78	4.22
(7) Born outside Canada	.33	6.47
(8) Born in Canada, outside Alberta	1.75	11.70
(9) Less than Grade 9	1.22	4.92
(10) Grades 9-13	-.40	3.50
(11) University	1.47	6.03
(12) Average Income, Male	1.74	7.09
(13) Average Income, Female	.94	4.12
(14) Male, Total	2.86	15.49
(15) Male, 0-14	.08	2.44
(16) Male, 15-19	-.03	2.68
(17) Male, 20-24	.74	2.99
(18) Male, 25-34	.46	2.36
(19) Male, 35-54	.07	4.18
(20) Male, 55-64	.52	3.38
(21) Male, 65 and over	.72	3.36
(22) Female, Total	-2.31	10.66
(23) Female, 0-14	.05	2.34
(24) Female, 15-19	2.73	19.93
(25) Female, 20-24	.89	4.71
(26) Female, 25-34	.63	2.36
(27) Female, 35-54	.14	3.01
(28) Female, 55-64	.20	2.25
(29) Female, 65 and over	1.02	4.14
(30) Single	.26	3.51
(31) Married	-.23	3.24
(32) English	-.66	5.21
(33) French	1.89	9.75
(34) Ukrainian	1.31	4.43
(35) German	1.97	8.11

4.3 RESEARCH METHODOLOGY

In the context of this research, the first step after data acquisition is to choose an appropriate number of classes. For 125 observations, the formula 5·log(n) (Equation 3.1) indicates that ten classes should not be exceeded. Therefore, nine class limits were determined by each method, for each variable. This limit may seem too large, but it helps to test the robustness of the classing systems. Since a minimum of five classes is often desirable, each method is also used to define four class limits for each variable.

The next step is to select classing methods. Eight classing systems are compared in this research: quantiles, equal steps, arithmetic progressions, geometric progressions, reciprocal progressions, nested-means, standard deviations and Jenks' "optimal" method. Fortran routines were written for each method to find a fixed number of class limits. These routines define class limits for thirty-five variables, and give the opportunity to test the eight classing methods. The classing routines are designed to ensure that classes do not overlap, and that they cover the full range of the data. Another Fortran routine is developed to measure how well class limits minimize variation within classes and maximize variation between classes, with three different indices: the Sum of Differences, the Tabular Accuracy Index, and the Goodness of Variance Fit. These three indices are calculated to measure the accuracy of each method and to determine the most appropriate classing system for each type of data distribution. The indices for each method are compared for a variable at a time. This indicates which method operates most efficiently for the type of frequency distribution examined. TAI and GVF have a common base of 1 and 100 for all variables. Therefore, the values of these two indices are compared across all the variables, one method at a time. This information determines which type of frequency distribution responds best with each method. The Sum of Differences index cannot be compared across variables, since it does not have a fixed base like the TAI or the GVF (e.g. base of 1 or 100). To determine to which type of frequency distribution the variable corresponds, histograms are plotted. Skewness and kurtosis indices are also calculated.

Ten variables are selected for a visual inspection of the different classifications. They reflect the different types of frequency distributions. These variables are: "Catholic," "Protestant," "Born in Canada, outside Alberta," "Less than Grade 9," "Grades 9-13," "University," "Male 0-14," "Female, Total," "Female 35-54," and "Single." The graphic representation of these variables includes the location of class limits and the area covered on the map by each class. A Postscript routine is used to plot a choropleth map of Edmonton, using the classes determined by a particular classing method. All this information is compiled and compared, and the efficiency of the different classing systems is discussed in the next chapter.

CHAPTER 5: EVALUATION OF CLASSIFICATION METHODS

In this chapter, the eight classing methods — quantiles, equal steps, arithmetic progressions, geometric progressions, reciprocal progressions, nested-means, standard deviations, and Jenks' "optimal" system — are tested with the thirty-five variables introduced in the last chapter. Differences for classifications with two different numbers of classes (five and ten) only seem to occur in the number of empty classes and the value of the TAI and GVF indices. Since class ranges are greater when a five class system is used, it is normal to observe fewer empty categories. A five class system offers a wider range of values of measure indices and these values seem to be lower than with a ten class system.

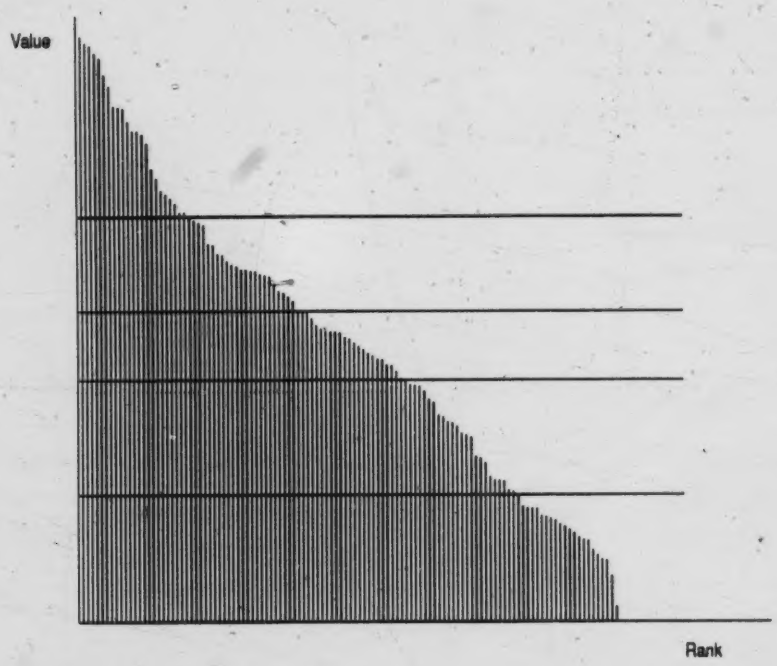
Since most of the classing methods do not consider the characteristics of the data distribution, most of the class limits they define may not minimize variation within classes and maximize variation between classes. The majority of the classing systems use a mathematical or statistical function. It may be possible to incorporate natural breaks in a classification defined by one of these systems. However, it would become hard to recognize which mathematical function is used. Furthermore, for the purpose of this study, class limits must remain intact. Nevertheless, to examine the different classifications visually, ten variables were selected. The eight classing systems were applied on each variable and histograms were created. Each histogram revealed the class limits, and the area covered on the map by each class. To visualize the way the different classing systems operate, each method is illustrated in this chapter with the variable that generated the best GVF. It includes a rank-size graph with the class limits (horizontal lines), and three other graphs: one showing the width of each class, another one showing the size (or frequency) of each class, and a last one showing the area covered on the map by each class. Area covered on the map is not really an issue in this study; however, it was of interest to observe this variable. To see if "good" classifications, according to the different measure indices, result in good choropleth maps, the best classification for each method was plotted as a choropleth map.

5.1 QUANTILES

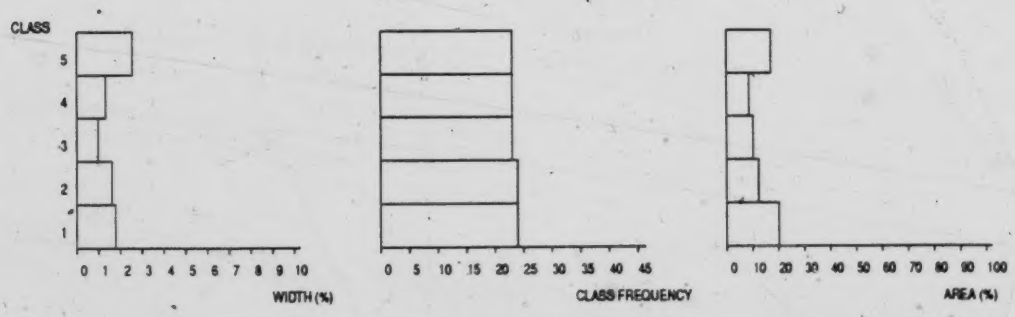
As indicated earlier, quantiles never show empty classes. Quantiles also ensure the maximization of information-content of a statistical map (Stegena & Csillag, 1987), since all classes are equal in frequencies (i.e. one class does not contain too much information, at the expense of the others).

According to the TAI and GVF measure indices, the reliability of the quantiles method may vary. The average GVF is relatively high: 91.923 for 10 classes and 82.463 for 5 classes. However, the ranges of the two measure indices are quite wide, so this method does not work well for all types of frequency distributions. It appears that quantiles should not be used with very skewed and, more importantly, highly peaked data distributions. Variables with the lowest TAI and GVF indices (i.e. "Canadian non-Albertan," "Male Total," "Female Total," "Female 15-19," and "French") show skewed (skewness greater than 1.5 or less than -2) and peaked (kurtosis greater than 9) data distributions. Conversely, the quantiles system operates most efficiently with a relatively normal (skewness=0.20) and platokurtic (kurtosis=2.25) frequency distribution: "Female 55-64" (Figure 5.1). Since classes have approximately equal frequencies, the graph of the area covered on the map by each class usually does not show major differences, unless the areal units vary considerably in size. Therefore, the resulting map should be relatively well areally balanced (Figure 5.2).

Figure 5.1 Female population between 55 and 64 years old (quantiles method)



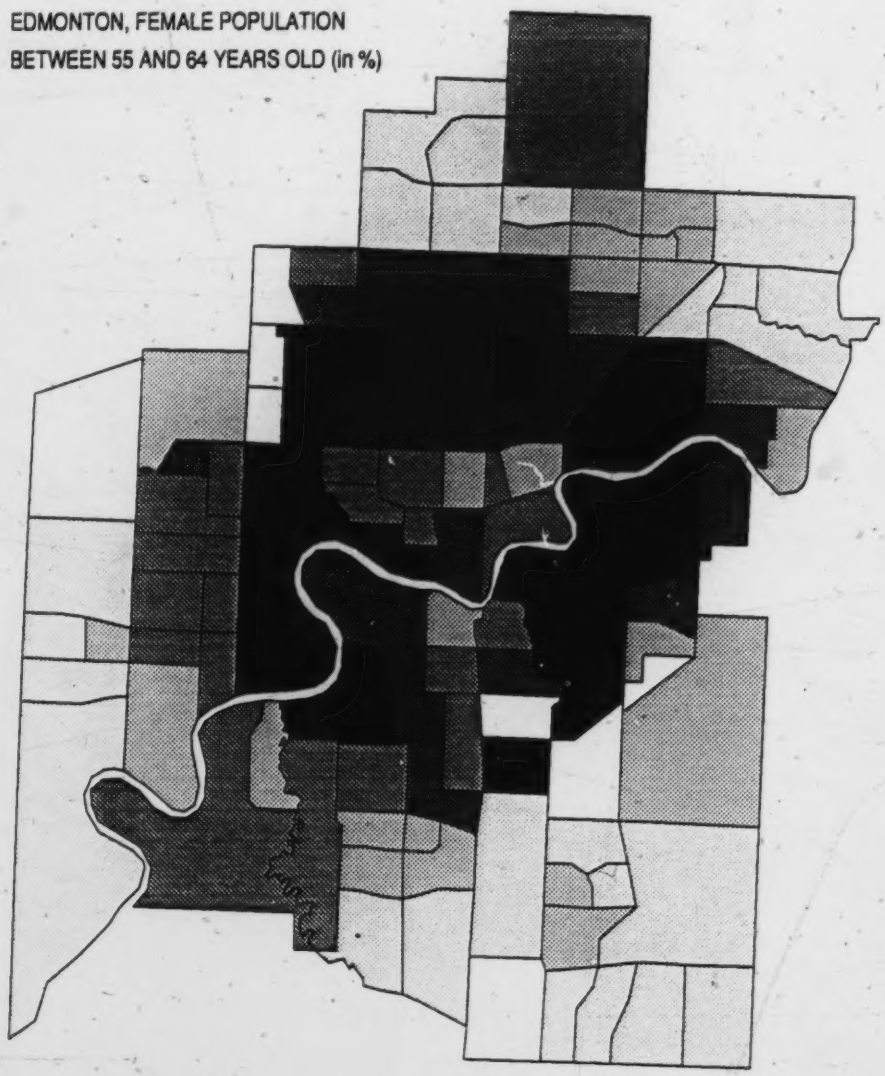
a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class



Source: Statistics Canada 1981

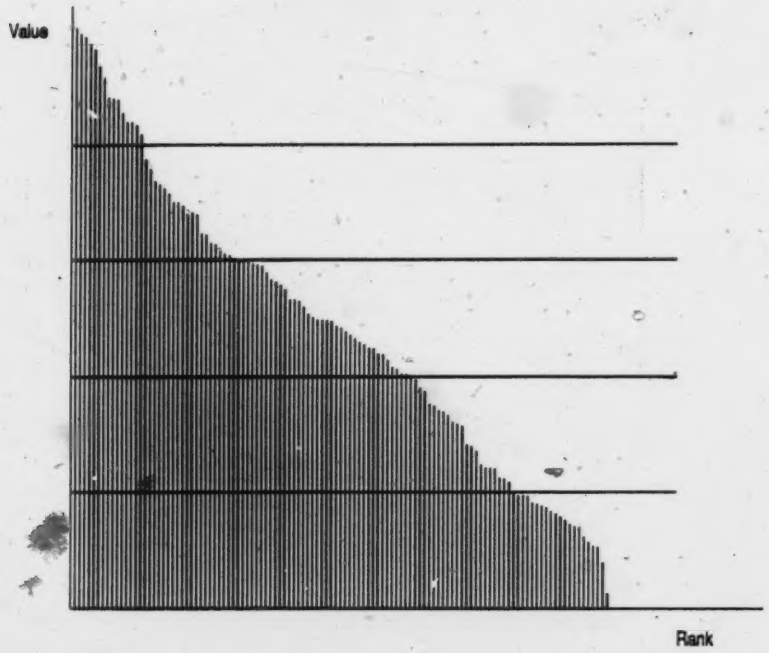
Figure 5.2 Choropleth map with classification defined by the quantiles system

5.2 EQUAL STEPS

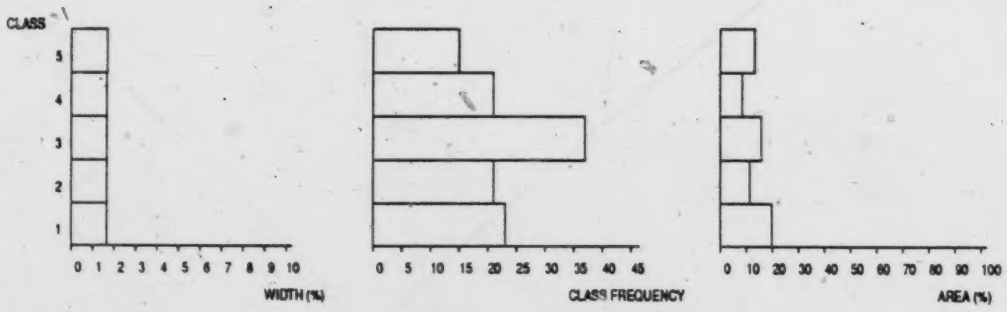
The equal steps technique has the drawback of allowing empty classes. When the data are divided into 10 categories, 17 variables show empty classes. Only 3 variables have this problem with 5 classes. A characteristic that emerges from these particular variables is that they tend to show a peaked frequency distribution. The degree of skewness does not appear to be a factor in the generation of empty classes.

In spite of the presence of empty classes, the equal steps method seems to give relatively good results, according to both GVF and TAI indices. With 10 classes, approximately 65% of the variables sampled get a TAI greater than 0.8. With 5 classes, this value is never attained. In both cases however, most of the variables get GVF values higher than 85. This method shows one of the smallest ranges in GVF indices for both 5 and 10 classes, which suggests that for such a sample the equal steps system appears to be relatively reliable. This observation appears to be in disagreement with Smith (1986), who considers the equal steps system unreliable because of its "little correspondence between classing accuracy and skewness" (Smith, 1986, p.64). The direction the data distribution takes, that is, normal, positively or negatively skewed, is not very significant. However, within the limits of the observed sample of variables, the best results appear with relatively normal distributions. The equal steps technique also seems to operate less effectively as the peakness of the distribution increases. This classing method is definitely more appropriate for platokurtic distributions (at best, flat distributions), and this agrees with Evans (1977). For example, the equal steps system gives the best classification with "Female 55-64," which is relatively normal and platokurtic (Figure 5.3). Both quantiles and equal steps methods operate very well with this variable characterized by a normal and platokurtic distribution. Therefore, it is not surprising that the resulting maps (Figures 5.2, 5.4) are quite similar. The main difference occurs with the central class.

Figure 5.3 Female population between 55 and 64 years old (equal steps method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

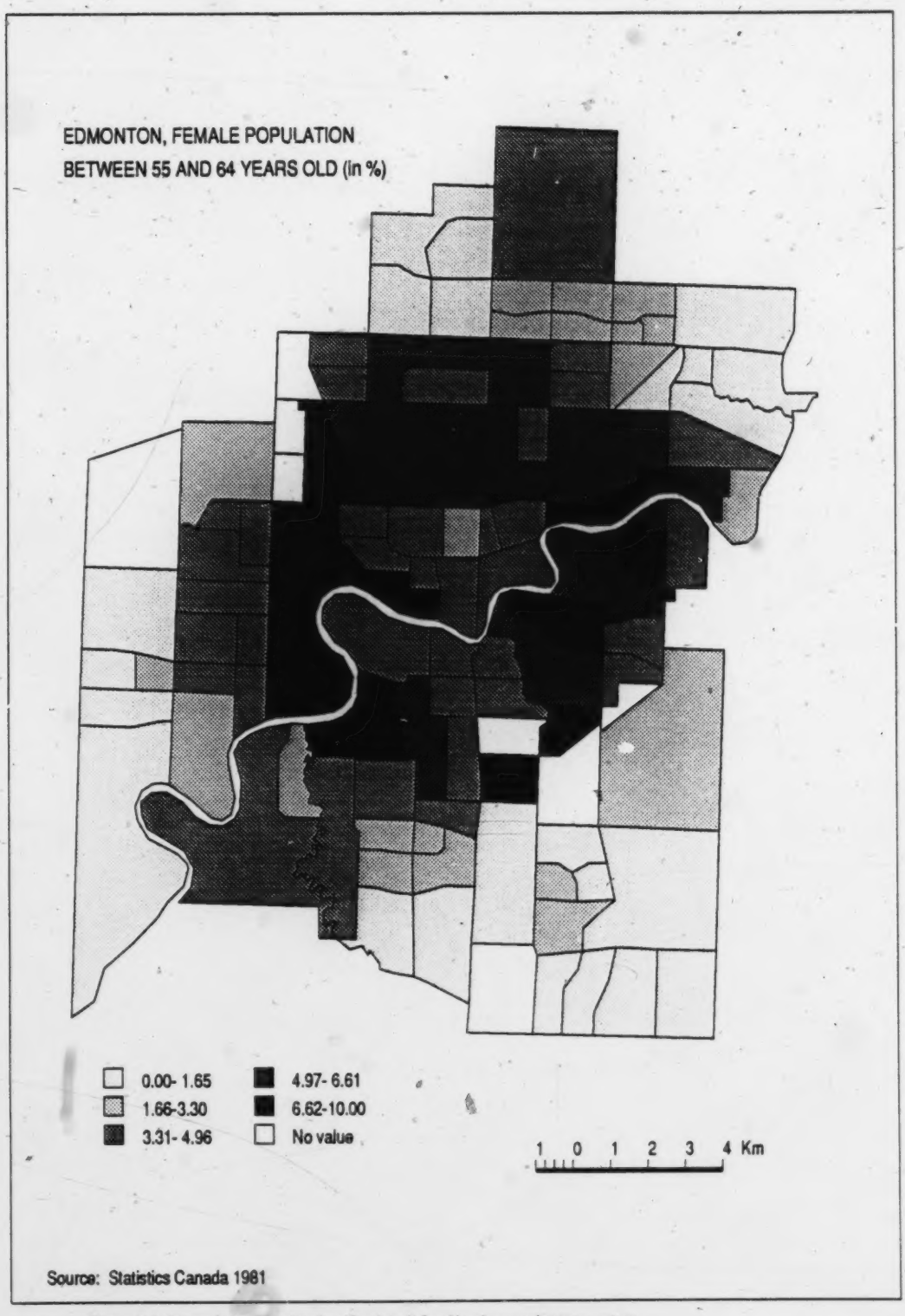


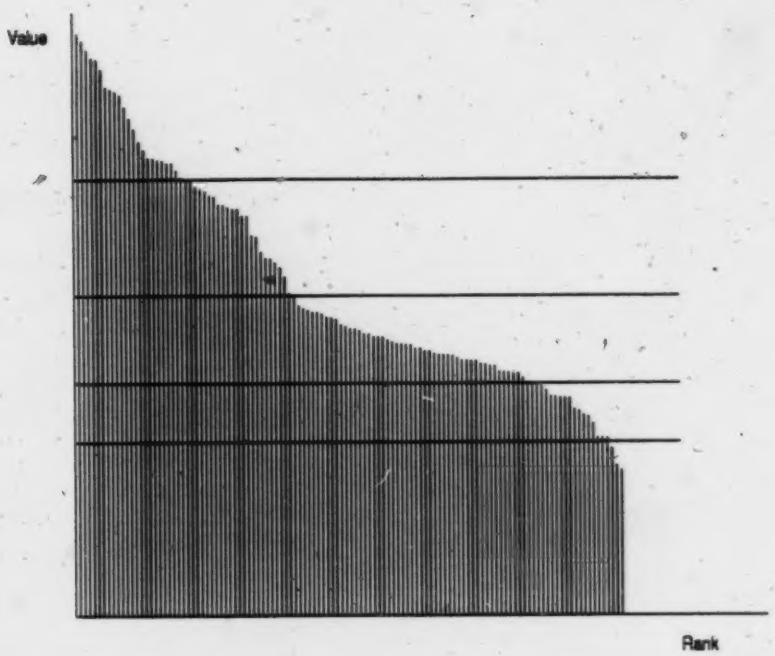
Figure 5.4 Choropleth map with classification defined by the equal steps system

5.3 ARITHMETIC PROGRESSIONS

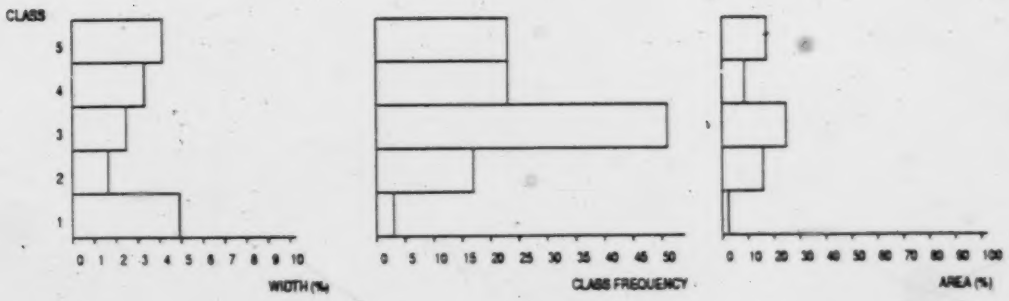
As with the preceding technique, the arithmetic progressions system has the drawback of allowing empty classes (22 with 10 categories; 8 with 5 categories). For the variables sampled, this method operates less effectively with a large number of classes. More than three empty classes can often be observed. A characteristic of these distributions is that they usually are leptokurtic. As with the equal steps technique, the direction of the skew does not seem to matter in the generation of unfilled categories.

In terms of GVF and TAI indices, arithmetic progressions could be considered as giving relatively good classifications. Most of the distributions have GVF indices greater than 85 and those indices show relatively small ranges for both 5 and 10 categories. The variables with the poorest classifications, according to the two indices, usually show peaked data distributions. Therefore, the arithmetic progressions method is least suitable for leptokurtic distributions. However, it does not seem possible to determine the kind of frequency distribution shape for which this system is most appropriate. The variables showing the highest GVF results with arithmetic progressions present all types of skews (positive, negative, normal) and kurtoses. Truran (1975) recommends this technique for small range distributions, and indeed, most of the variables with the highest GVF index show a range of values smaller than 20%. However, since variables with lower results do not necessarily have a wide range of values, arithmetic progressions should not be systematically used with data distributions offering a small range of values. Figure 5.5 shows that arithmetic progressions tend to have narrow classes towards the lower end of the frequency distribution, Therefore, the resulting map shows less of the light tones.

Figure 5.5 Female population between 25 and 34 years old (arithmetic progressions method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

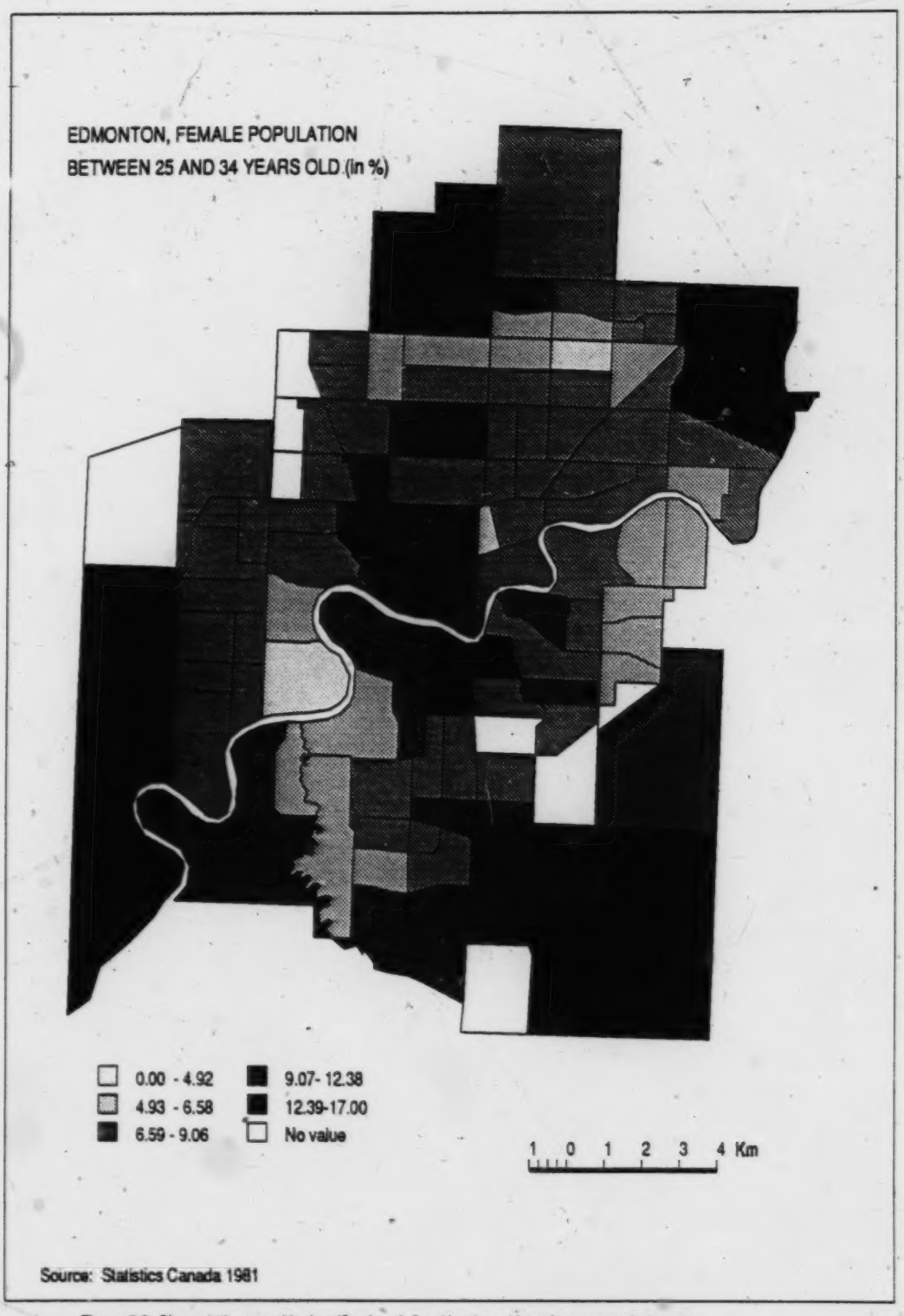


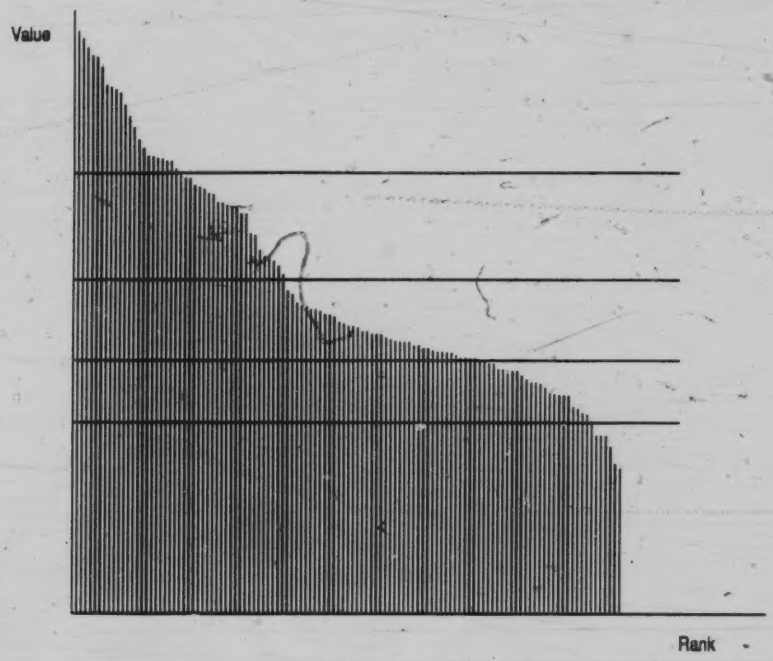
Figure 5.6 Choropleth map with classification defined by the arithmetic progressions system

5.4 GEOMETRIC PROGRESSIONS

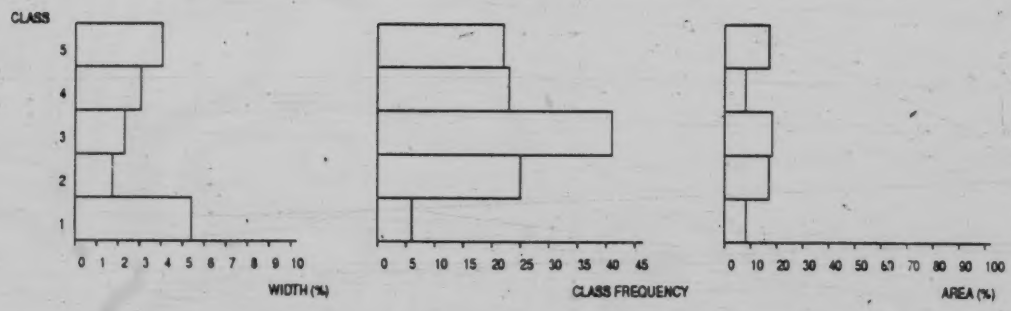
Since the logarithm of zero is undefined, no classification can be achieved by the geometric progressions system when the minimum value of a data distribution equals zero. Therefore, variables with a minimum value of zero have been ignored, and only 19 variables are classed by geometric progressions. These geometrically classified variables are: "Density," "Catholic," "Protestant," "Born in Canada," "Born in Alberta," "Grades 9-13," "Average Income, Male," "Average Income, Female," "Male, Total," "Male 0-14," "Male 20-24," "Male 25-34," "Male 55-64," "Female, Total," "Female 0-14," "Female 25-34," "Single," "Married," "English." Geometric progressions can produce empty categories. This system seems particularly sensitive to severely leptokurtic distributions, since as many as 7 or 8 empty classes may be observed. This technique also does not operate effectively with negatively skewed data distributions. This is obvious, since the equation employed is supposed to be used with positively skewed distributions.

Geometric progressions show results comparable to other methods. The average GVF value is close to the highest ones observed with other techniques (83.76 with 5 classes; 94.284 with 10 classes). This system operates more effectively for positively skewed data distributions. Figure 5.7 gives an example of a relatively good classification built with the geometric progressions systems. As expected, this system operates quite similarly to the arithmetic progressions method and the maps for both systems differ only slightly. However, a relatively wide range of values within the data distribution can lead to a relatively good classification even if the skew is not positive. Therefore, regardless of the skew, if the frequency distribution is very peaked, this system may not work as well as with a relatively normal distribution.

Figure 5.7 Female population between 25 and 34 years old (geometric progressions method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

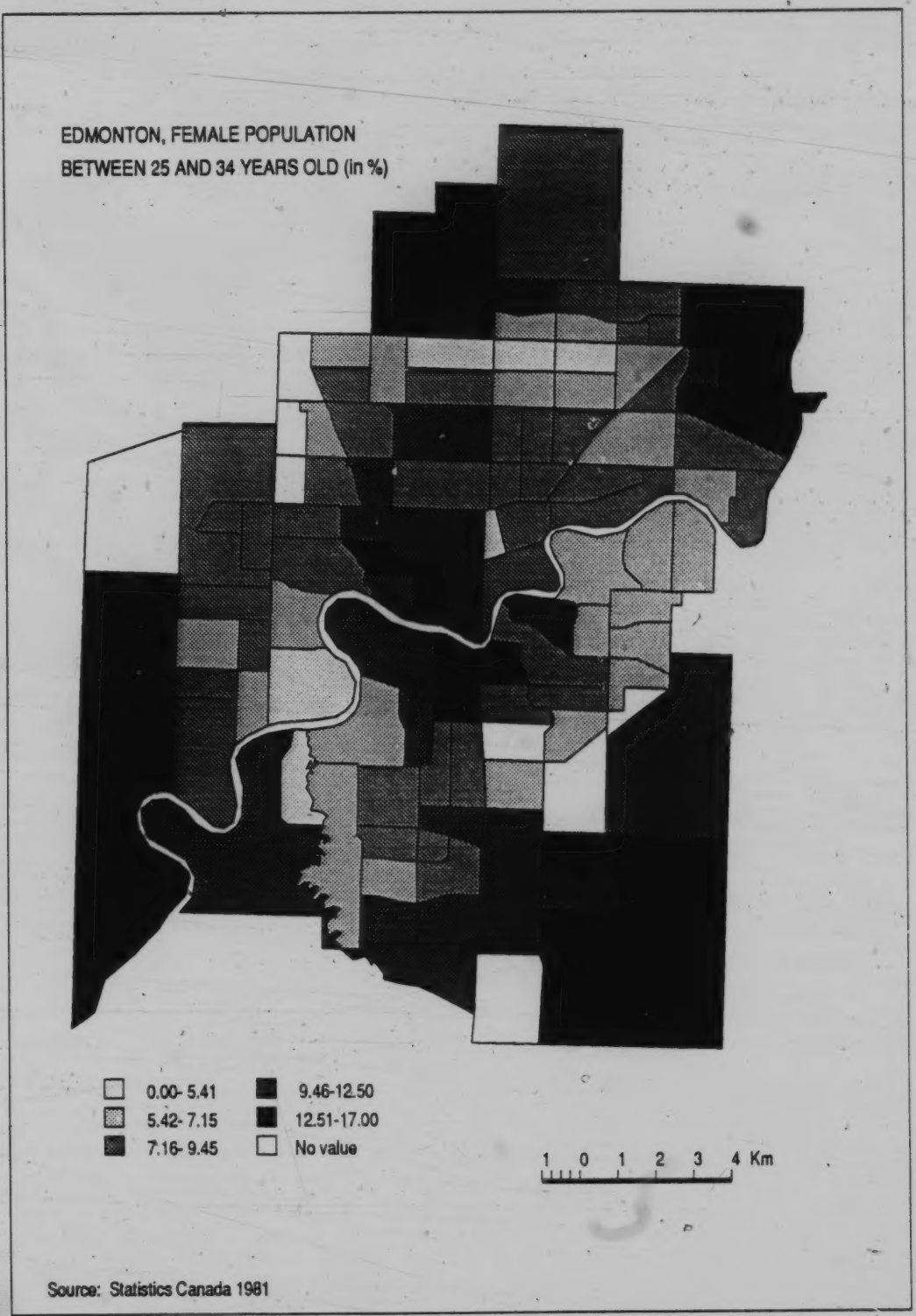


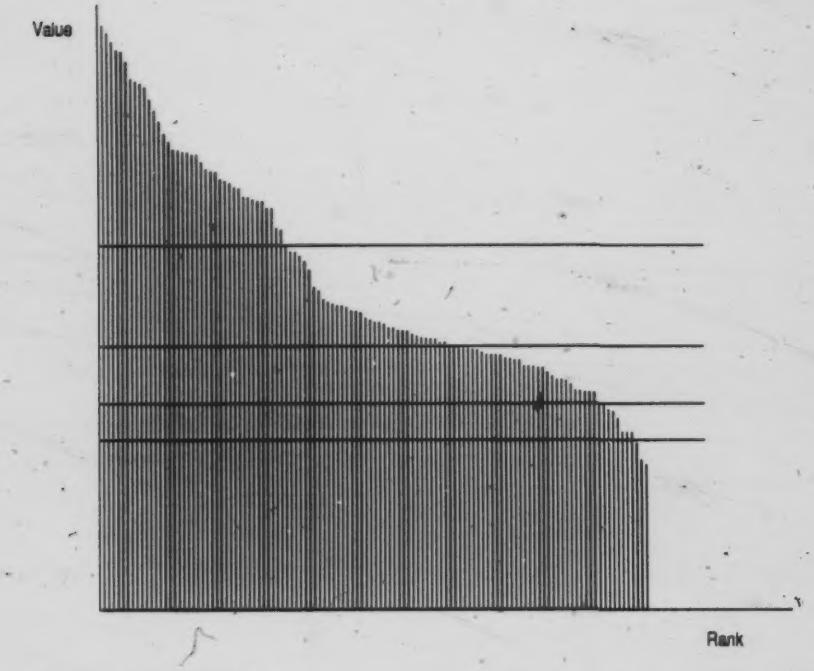
Figure 5.8 Choropleth map with classification defined by the geometric progressions system

5.5 RECIPROCAL PROGRESSIONS

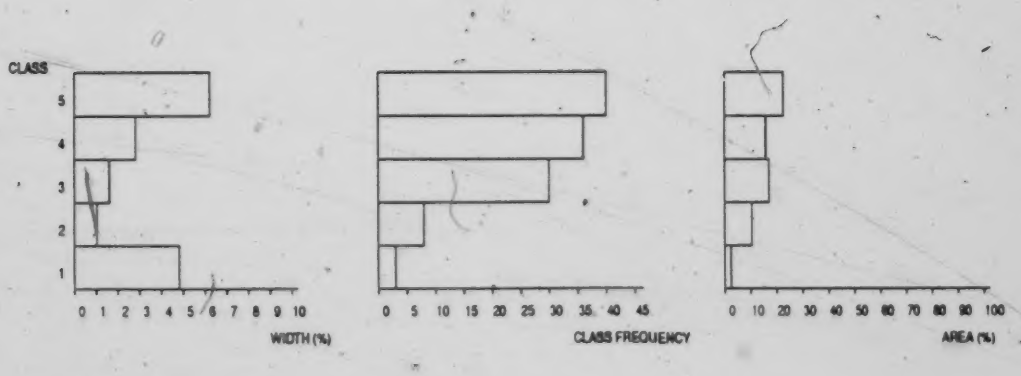
Since the formula for reciprocal progressions implies the division of one by the minimum value of a distribution, when this minimum is zero, no results can be obtained. Therefore, class limits are calculated for only nineteen variables (named in the preceding section). The reciprocal method may also show empty classes, and these tend to occur when a data distribution is relatively peaked.

The reciprocal method may seem unreliable on the basis of GVF values. These values range from 44.653 to 92.342 with 5 classes and from 64.811 to 97.933 with 10 classes ("Average Income Female" (lowest); "Female 25-34" (highest)), and the TAI can be as low as .307 ("Average Income Female"). Therefore, even with a GVF index average of 90.194 with 10 classes, reciprocal progressions do not give the best overall classing results. The variables with the worst classing results tend to have peaked data distributions. The variable with the best classification, "Female 25-34," is positively skewed and relatively flat (Figure 5.9). This technique shows a definite progression in the definition of class limits, and higher classes have high frequencies. Therefore, it is not surprising to observe very dark maps, when classed with reciprocal progressions (Figure 5.10). Surprisingly, reciprocal progressions tend to work with negatively skewed distributions as well as with positively skewed ones. Davis (1974) and Paslawski (1984) recommended that this system be employed with very large positively skewed data distributions, but based on the results on this research, it is difficult to agree with these theoretical statements.

Figure 5.9 Female population between 25 and 34 years old (reciprocal progressions method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

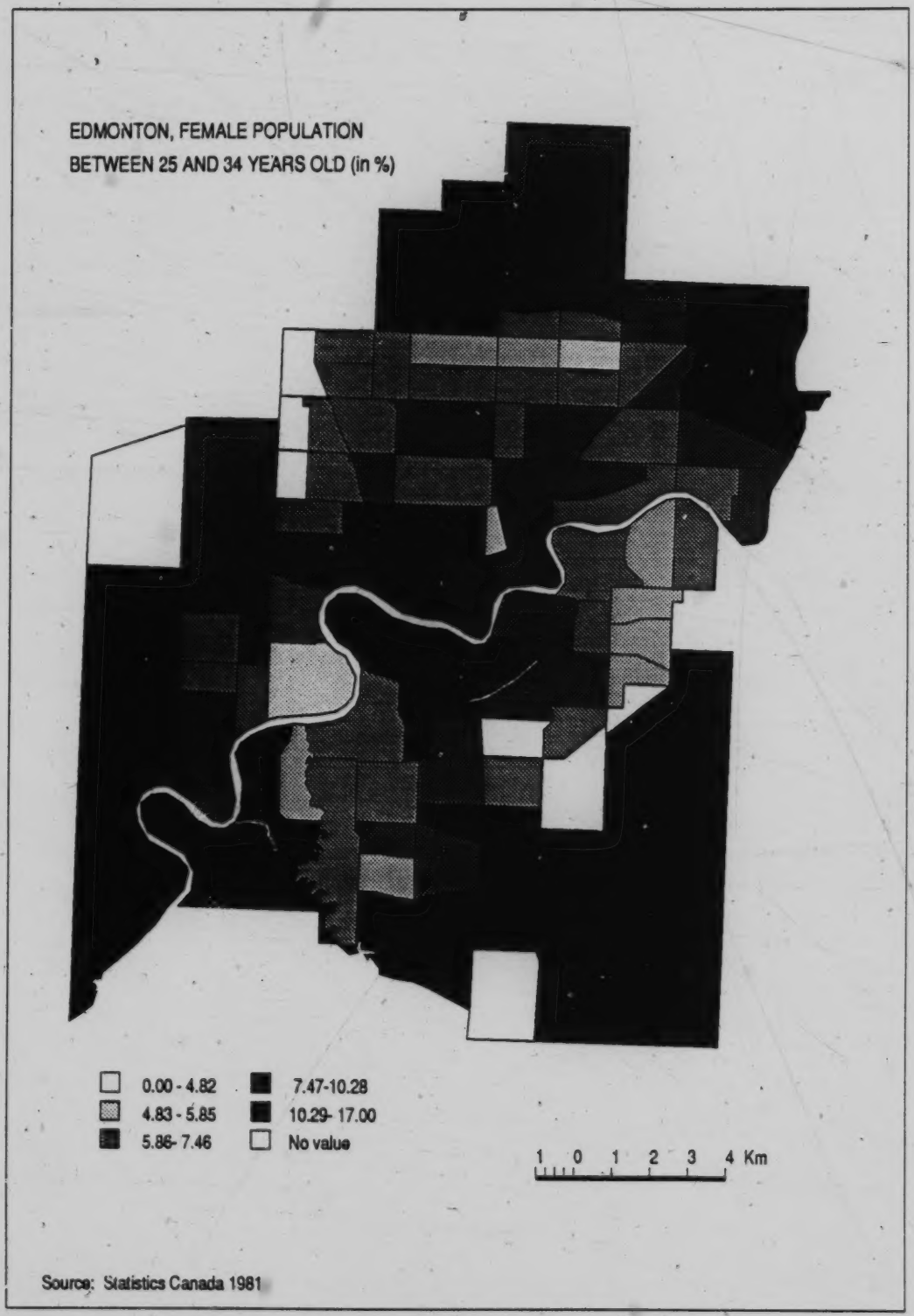


Figure 5.10 Choropleth map with classification defined by the reciprocal progressions system

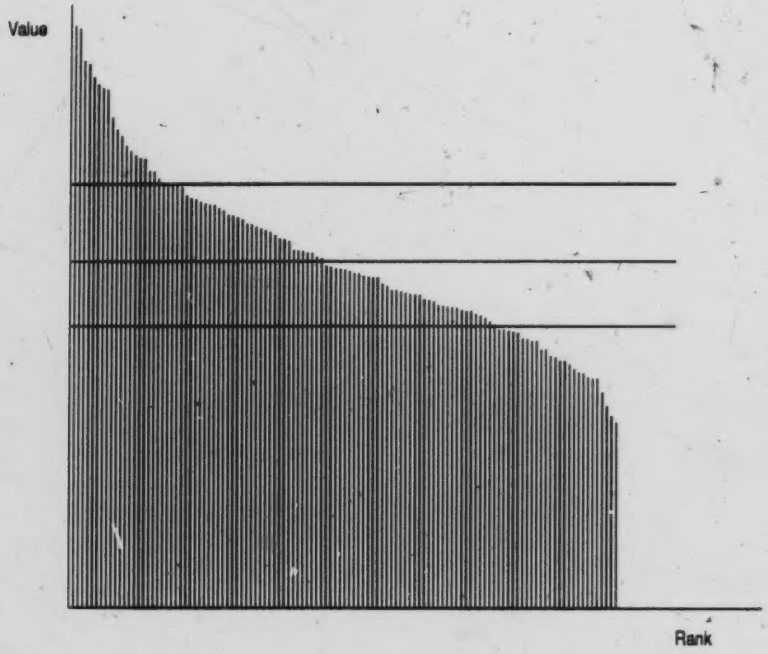
5.6 NESTED-MEANS

The nested-means technique has a very low risk of generating empty classes, unless the number of classes is too large for the data distribution. None of the variables sampled showed empty categories with a nested-means classification. Since this method always employs the means of a set of data in a distribution, there is at least one value on each side of the mean, unless there is only one value in this data set. As Chang (1974) and Paslawski (1984) indicated, and according to GVF and TAI indices, nested-means appear to be more appropriate to data distributions which are close to normal, than to skewed or leptokurtic distributions. Most of the variables show GVF indices higher than 80, and variables with the highest scores usually show normal or approximately normal data distributions. For example, the best classification is given for a normal and peaked distribution: "Male 35-54" (Figure 5.11). Because of the way this technique operates (it employs means), classes tend not to differ significantly in size and tend to cover approximately the same amount of surface on the map (Figure 5.12). Finally, a marked peak does not seem to have a critical influence.

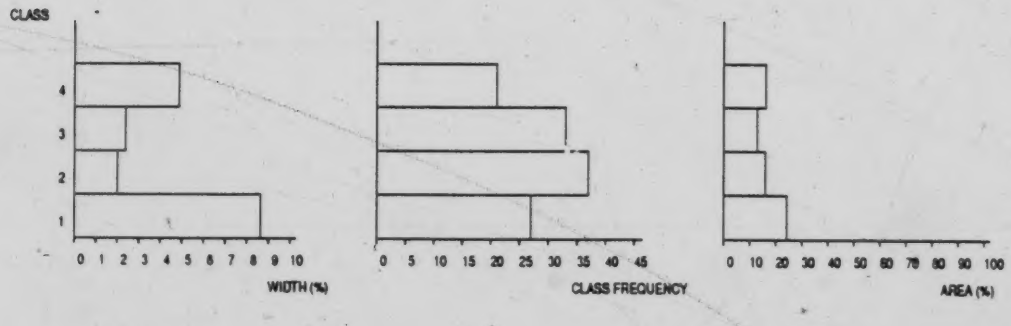
5.7 STANDARD DEVIATIONS

The standard deviations technique may create empty classes for a data distribution. Unfilled categories tend to occur with leptokurtic distributions. Nevertheless, the standard deviations technique does have its advantages. For more than half of the variables, this technique proposed classes with a GVF index superior to 85. However, this index can reach values as low as 56.747 (5 classes) or 73.367 (10 classes), both with the "Average Income Female" variable. Consequently, this technique gets average GVF values of 84.281 and 91.2 with 5 and 10 categories respectively. It may not be possible to determine what kind of data distribution is less appropriate for a standard deviations classification, because of the

Figure 5.11 Male population between 35 and 54 years old (nested means method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

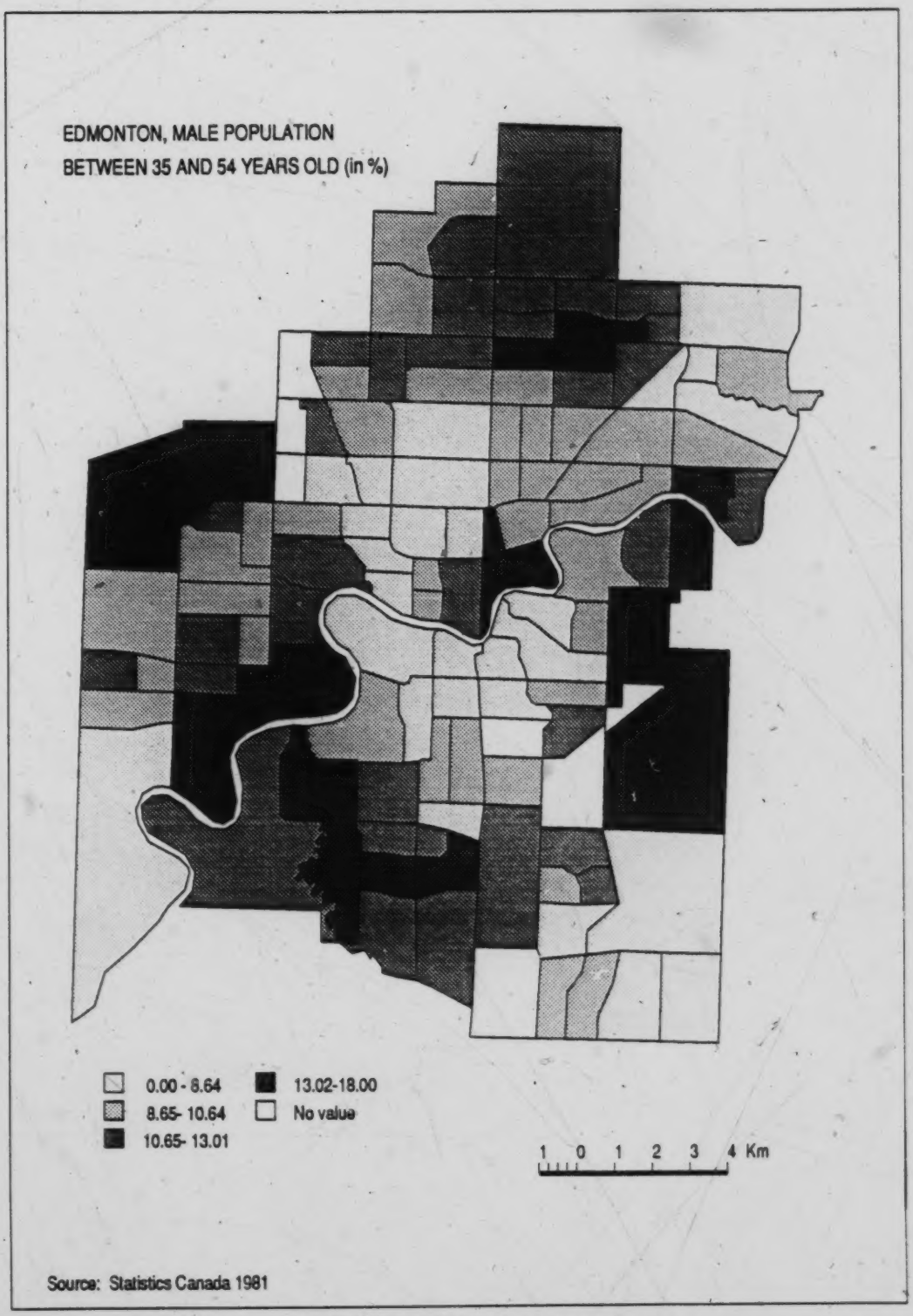


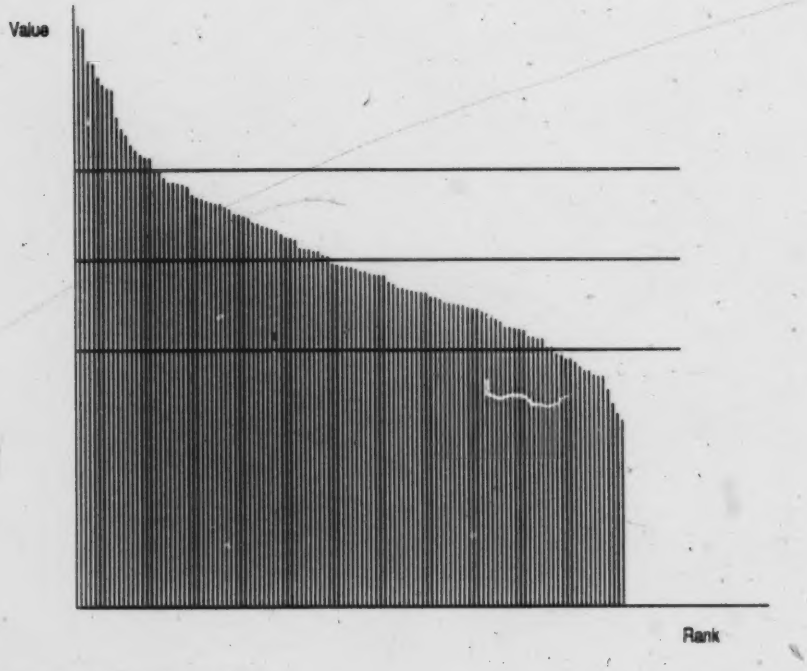
Figure 5.12 Choropleth map with classification defined by the nested-means system

small number of lower GVF or TAI indices. It still appears that this technique is more suited for approximately normal distributions (e.g. "Male 35-54," Figure 5.13).

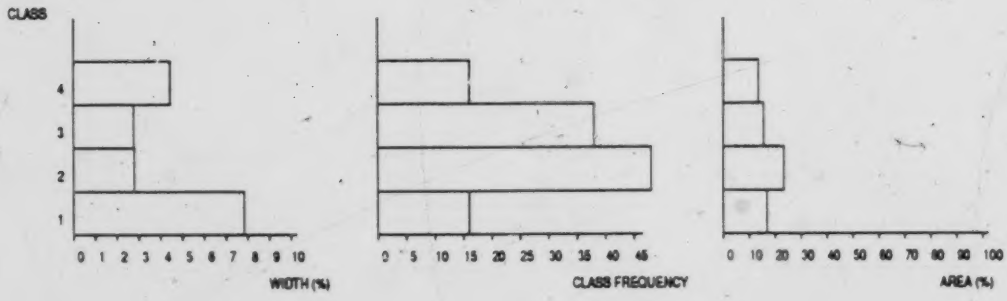
5.8 "OPTIMAL"

The "optimal" method, either using variance or absolute deviations, may give the best classifications, according to the indices. In most cases, GVF indices are superior to 90, and in the case of a 10-class map, most of the GVF values are greater than 97. Since they base their class limit determination on absolute deviations or on variance, both types of "optimal" methods almost always show the best results according to the TAI and the GVF indices; TAI is calculated from absolute deviations and GVF is derived from variances. This explains why TAI generally produces higher values for the "optimal" method from absolute deviations, while GVF shows higher values for the "optimal" routine from variances. This method becomes inefficient when a data distribution seems to be unimodal, since it tries to identify natural groups to find class limits. In the context of this research, a GVF index less than 97 with 10 classes, or less than 80 with 5 classes, indicates a problem: in such cases, the distribution may be unimodal or only have two slight modes. When this occurs (with variables such as "Male 0-14," "Male 15-19" or "Female 0-14"), the GVF index can be as low as 18 or 19, and such values considerably lower the GVF average for this method. This technique appears to be most effective when the unimodal data distributions are excluded, and generally the type of distribution does not seem to be significant. The "optimal" system operates as well with normal or skewed distributions. Nevertheless, an additional problem remains. The memory space required to accommodate the matrices generated by this routine precludes the classification of the Average Income variables.

Figure 5.13 Male population between 35 and 54 years old (standard deviations method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

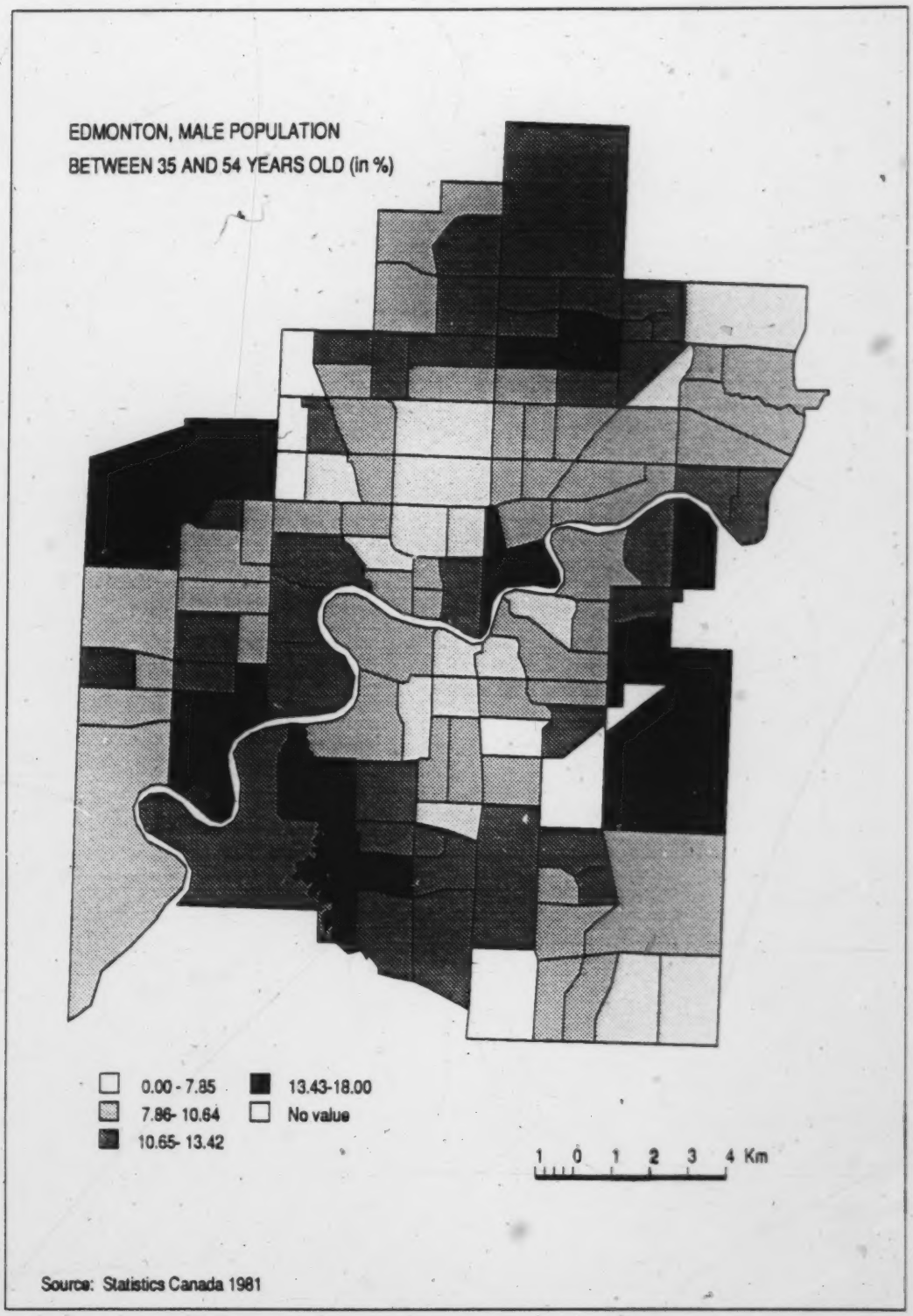
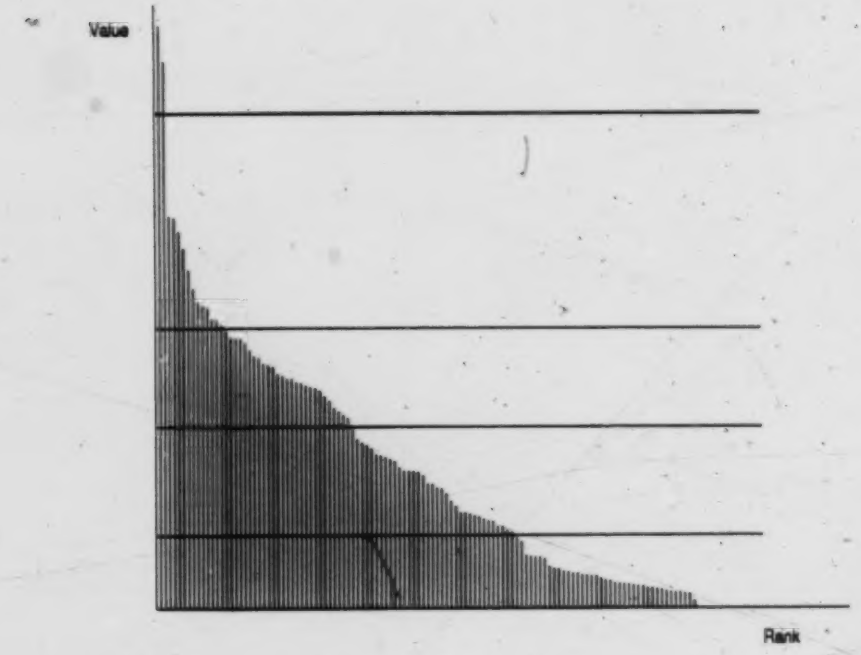
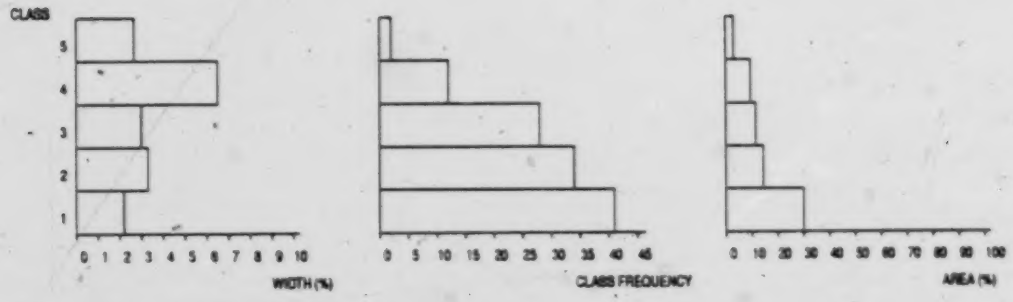


Figure 5.14 Choropleth map with classification defined by the standard deviations system

Figure 5.15 Female population 65 years old and over (Jenks' "optimal" method)



a. Rank-size graph with class limits



b. Classes width

c. Classes frequency

d. Area covered by each class

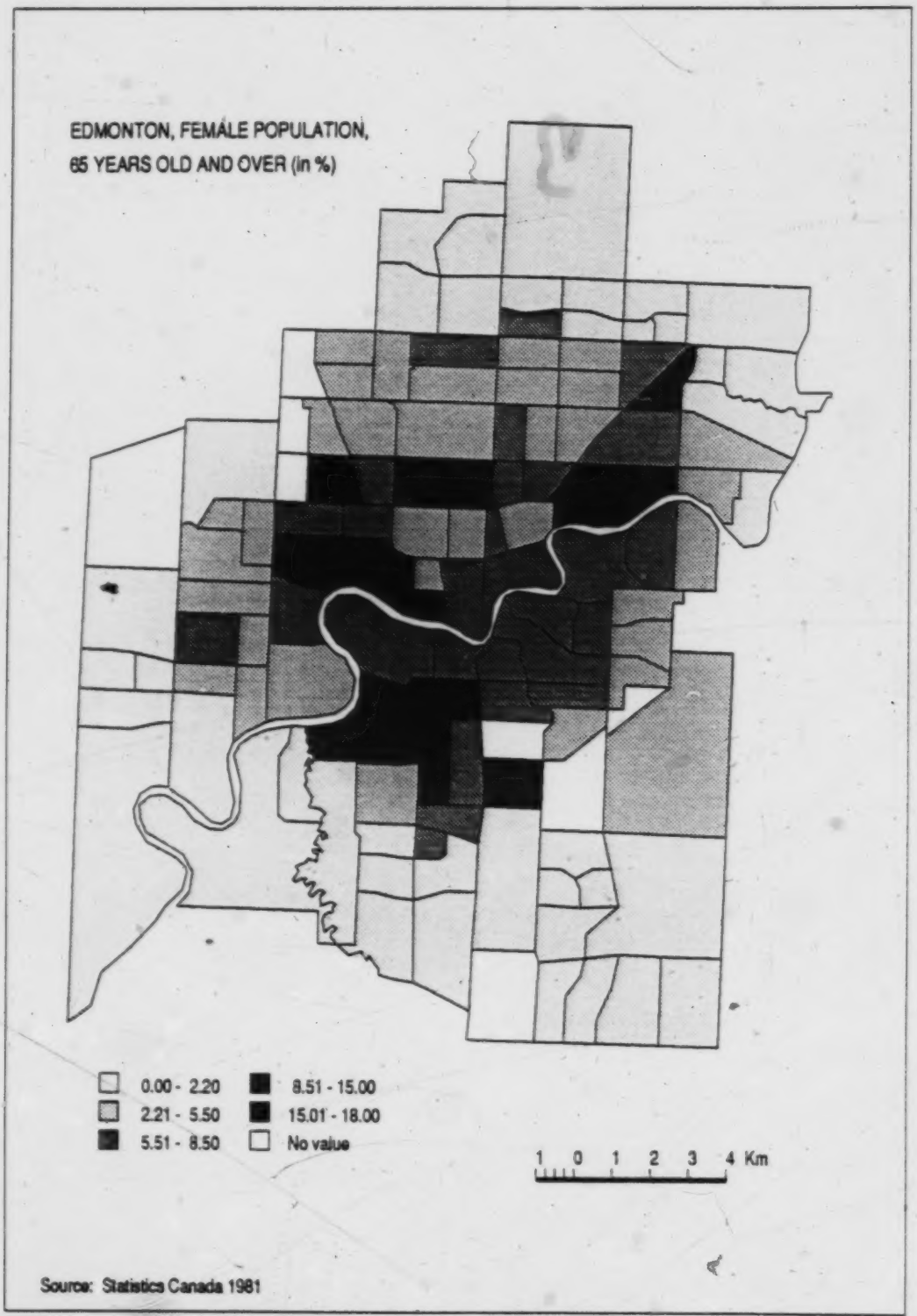


Figure 5.16 Choropleth map with classification defined by the "optimal" system

5.9 CONCLUDING COMMENTS

Evans (1977) stated that, depending on the shape of the data distribution, some methods are more appropriate than others to define class limits. The nested-means method may be used with a normal distribution, geometric progressions may be used with very skewed distributions, and arithmetic progressions may be useful for small range distributions. Otherwise, when the kind of data distribution is hard to define or seems hard to divide, the quantiles method may be the most appropriate (Davis, 1974).

This research found that for data distributions with skewness equal or close to zero (normal distribution), the equal steps method seems to be the appropriate option. Both TAI and GVF tend to show higher values with the equal steps method than with the other techniques when the data distribution is close to normal. The literature (Chang, 1974; Paslawski, 1984) recommends the standard deviation method to divide a normal distribution into classes. This technique is only a second option after the comparison of the eight methods. Equal steps also seems a good alternative for data distributions that are only slightly skewed (skewness less than 1 or greater than -1). The Sum of Differences index selects nested-means or standard deviations methods for slightly skewed distributions (close to normal), and this supports statements from Paslawski (1984) and Chang (1974). Although Evans (1977) rejects the arithmetic method, for the variable sampled, the arithmetic technique tends to produce relatively good results for very positively skewed distributions. GVF and TAI indices select the standard deviations or the nested-means method for distributions with a skewness greater than 2 or smaller than -2. For a flat distribution, the equal steps method appears to be the most appropriate choice.

As expected, different methods operate more efficiently with a particular type of frequency distribution. Quantiles and equal steps methods operate best with a relatively normal distribution (e.g., "Female, 55-64"). Arithmetic, geometric, and reciprocal progressions give their best results with the variable "Female, 25-34," which is slightly positively skewed. Nested-means and standard deviations systems work the most efficiently

with a relatively normal data distribution, "Male, 35-54." Jenks' "optimal" method does not depend on the shape of the data distribution to operate efficiently, but on the presence of modalities in the distribution. Therefore, for a normal distribution, a cartographer may choose between quantiles, equal steps, nested-means, and standard deviations methods, to define class limits. The cartographer may also have the choice between arithmetic, geometric, and reciprocal progressions to define class limits for a skewed data distribution.

Finally, it appears that the Sum of Differences index, tends to select the "best" classification in agreement with theoretical concepts, whether or not the class limits correspond with natural breaks. Conversely, since they measure absolute deviations or variance for every class, TAI and GVF tend to favour class limits which reflect most the natural groups of a distribution.

CHAPTER 6: CONCLUSION

This research examined the classification process in choropleth mapping. Eight different classing systems were tested with thirty-five variables to examine if a relation between classing methods and type of frequency distribution exists. It appears that there *is* such a relationship. Four systems operate more efficiently with normal distributions: quantiles, equal steps, nested-means, and standard deviations. Arithmetic, geometric and reciprocal progressions are more appropriate for skewed data distributions. However, none of these systems produce good class limits with pronounced leptokurtic distributions.

This study essentially considered the statistical aspect of classification when evaluating each method. However, areal balance must be considered in the creation of a choropleth map. It is desired to create a relatively balanced map, where the different classes all tend to occupy approximately equal aggregate geographic areas. Because it considers the number of areal units by class, the quantiles tends to produce maps which are relatively well balanced. However, this system does not consistently guarantee an areally balanced map. The nested-means system also tends to give relatively well balanced maps. In genera the quantiles and nested-means systems offer classifications with the least variations in area covered by each category. The classing methods examined do not consider areas, but only the value of the variable for each mapping unit. Therefore, one cannot expect to achieve a classification with a relatively good balance in the areas covered by each class. To improve the appearance of the map, the cartographer may choose shades that will try to mask the areal imbalance of the map. It also can be possible to use the clinographic curve or the frequency cumulative curve to obtain a better areally balanced map.

Considering the results of this research, it seems possible to develop an automated approach to class a data distribution. Given a computer system with adequate memory and processing speed, a form of univariate clustering, using eigenvalues (as discussed in Chapter 3) can be employed to define the number of classes. However, this technique considers only the statistical aspect of the data and ignores factors such as the purpose and

the type of audience of the map. When classing data automatically, the user may be presented with different options, which depend on the size of the sample and the computer capacity. If the frequency data distribution is multimodal, the "optimal" method appears to be the most effective classing technique. However, the problem of non-contiguous class limits still remains. This can easily be solved with an automated routine which finds an intermediate value between two neighbouring class limits. This value would become the new class limit between two adjacent classes. If the "optimal" technique cannot be used, a routine (ensuring that there is no empty class) can automatically select one or more classing methods depending on the type of distribution. Class limits can be evaluated and the best classification is chosen for the map. It therefore seems possible to create an automated classification routine which defines good class limits without the intervention of a user.

In conclusion, there is adequate evidence that there are regularities in the process of classification in choropleth mapping and that a classification model could be integrated in an automated system, producing maps that are cartographically sound.

6.1 TOWARDS THE AUTOMATION OF CHOROPLETH MAPPING?

In Chapter 2, it was argued that classification was one of the four principal procedures — preliminary decisions, classification, polygon fill and finalizing — in choropleth mapping. The conclusions of this study are extremely encouraging: it is clear that classification procedures readily lend themselves to automation. Certain types of distributions are evidently more amenable to classification by some methods than by others. This forms the basis of a model of classification in choropleth mapping. Such a model could lead to the creation of an integrated choropleth mapping programme.

It remains to investigate whether the other three procedures can be formalized, and eventually automated. At this point in time they still seem to require the intervention of the cartographer (or user, if in the context of a computer routine). This is clearly an exciting

area for possible research in the future where all the regularities of choropleth map design must be determined. In the event of the complete automation of choropleth mapping procedures, the resulting routine would follow this certain order to operate. Knowing the motivation factors of the map, it would only take a few minutes to define classes automatically. With these classes, a routine would automatically fill the polygons. If a colour fill is employed, this routine should consider more than just the three physical characteristics of colour (hue, saturation and value), but also the perceived intensity or value of colour. The last set of routines must ensure that all the elements on the map have an appropriate size and location, to ensure a cartographically accurate but esthetically pleasing choropleth map. This seems to be the step where most of the difficulties in achieving a computer routine would occur. The different elements of information (e.g., title, legend, scale) should not look more important than the principal figure.

If all the procedures to create a choropleth map can be carried out objectively, and put into a model, this can be done for other types of maps as well. The creation of a completely automated choropleth mapping software could be a first step for the generation of a mapping expert system (Gondran, 1986). The number of human operations or procedures transposed into computer routines is increasing rapidly, and cartography is not excluded. It is therefore our duty, as cartographers, to ensure that these automated procedures are giving a product as good, if not better, as the product done manually.

Bibliography

Armstrong, R.W. 1969. Standardized Class Intervals and Rate Computation in Statistical Maps of Mortality, Annals of the Association of American Geographers, 59(2):382-390.

Barber, Gerald M. 1988. Elementary Statistics for Geographers, The Guilford Press, New York, 513 pages.

Brassel, Kurt E. and Utano, Jack J. 1979. Design Strategies for Continuous-tone Area Mapping, The American Cartographer, 6(1):39-50.

Burrough, P.A. 1986. Principles of Geographical Information Systems for Land Resources, Clarendon Press, Oxford, 194 pages.

Caldwell, Douglas R. 1981. Design Principles and Automated Choropleth Mapping, How to Design an Effective Graphics Presentation, Harvard Library of Computer Graphics:5-6.

Chang, Kang-Tsung. 1974. An Instructional Computer Program on Statistical Class Intervals, The Canadian Cartographer, 11(1):69-77.

Chang, Kang-Tsung. 1978. Visual Aspects of Class Intervals in Choropleth Mapping, The Cartographic Journal, 15(1):42-48.

Chang, Kang-Tsung. 1979. Class Intervals in Choropleth Mapping, Thematic Map Design, Harvard Library of Computer Graphics: 5.

Coulson, Michael R.C. 1987. In the Matter of Class Intervals for Choropleth Maps: With Particular Reference to the Work of George F. Jenks, Cartographica, 24(2):16-39.

Cuenin, René. 1972. Cartographie Générale, Notions Générales et Principes d'Elaboration, Tome 1, Editions Eyrolles, Paris, 324 pages.

Cuff, David J. and Mattson, Mark T. 1982. Thematic Maps; Their Design and Production, Methuen, New York, 169 pages.

Davis, Peter. 1974. Science in Geography 3; Data Description and Presentation, Oxford University Press, Oxford, G.B., 119 pages.

DeBrommer, S. 1969. Vers une Cartographie Moderne; Evolution des Buts et des Aspects, Bulletin du Comité Français de Cartographie, 4(42):250-258.

Dickinson, Gordon Cawood. 1963. Statistical Mapping and the Presentation of Statistics, Edward Arnold, London, 160 pages.

Ebdon, David. 1985. Statistics in Geography, Second Edition, Basil Blackwell, Oxford, G.B., 232 pages.

Evans, Ian S. 1977. The Selection of Class Intervals, Contemporary Cartography; Transactions New Series 2(1):98-124.

Fisher, Howard T. 1979. Thematic Cartography (Paper Number One), Thematic Map Design, Harvard University:8-37.

Fisher, Howard T. 1982. Mapping Information; The Graphic Display of Quantitative Information, Abt Books, Cambridge, Mass., 384 pages.

Gondran, Michel. 1986. Introduction aux Systèmes Experts, Editions Eyrolles, Paris, 100 pages.

Jenks, George F. 1976. Contemporary Statistical Maps — Evidence of Spatial and Graphic Ignorance, The American Cartographer, 3(1):11-19.

Jenks, George F. 1977. Optimal Data Classification for Choropleth Maps, Department of Geography Occasional Paper 2, The University of Kansas, Lawrence, Kansas, 25 pages.

Jenks, George F. and Caspall, F. 1971. Error and Choroplethic Maps: Definitions, Measurement, Reduction, Annals of the Association of American Geographers, 61(2):217-224.

Jenks, George F. and Coulson, M.R. 1963. Class Intervals for Statistical Maps, International Yearbook of Cartography, 3:119-134.

Johnston, R.J. 1968. Choice in Classification: the Subjectivity of Objective Methods, Annals of the Association of American Geographers, 58(3):575-589.

Kretschmer, Ingrid. 1978. Les Problèmes Urgents de la Cartographie Théorique, Bulletin du Comité Français de Cartographie, 1(75):13-16.

Lai, Poh-Chin. 1986. RANDMAP/Rand McNally & Company (software review), The American Cartographer, 13(4):355-358.

Lavin, Stephen & Archer, J.Clark. 1984. Computer-produced Unclassed Bivariate Choropleth Maps, The American Cartographer, 11(1):49-57.

Lescort, Adrien. 1985. Intelligence Artificielle et Systèmes Experts, Editions Cedic/Nathan, Paris, 141 pages.

Maceachren, Alan M. 1985. Accuracy of Thematic Maps/ Implications of Choropleth Symbolization, Cartographica, 22(1):38-58.

Mackaness, W.A. 1986. Towards a Cartographic Expert System, Auto-Carto, London, G.B.:578-587.

Marles, A.C. 1984. Identifying and Meeting Map User Needs, Cartographica, 21(1):135-138.

Monkhouse, F.J. & Wilkinson, H.R. 1971. Maps and Diagrams: Their Compilation and Construction, Methuen & Co, London, G.B., 522 pages.

Monmonier, Mark Stephen, 1973. Analogs Between Class-Interval Selection and Location-Allocation Models, The Canadian Cartographer, 10(2):123-131.

Monmonier, Mark Stephen. 1973. Eigenvalues and Principal Components: A Method for Detecting Natural Breaks for Choroplethic Maps, American Congress on Surveying and Mapping, Proceedings: 252-264.

Monmonier, Mark Stephen. 1975. Class Intervals to Enhance the Visual Correlation of Choroplethic Maps, The Canadian Cartographer, 12(2):161-178.

Monmonier, Mark Stephen. 1978. Modifications of the Choropleth Technique to Communicate Correlation, International Yearbook of Cartography, 18:143-158.

Monmonier, Mark Stephen. 1982. Flat Laxity, Optimization, and Rounding in the Selection of Class Intervals, Cartographica, 19(1):16-27.

Morrison, Joel L. 1975. Map Generalization: Theory, Practice and Economics, Auto-Carto II: 99-112.

Muller, Jean-Claude. 1976. Numbers of Classes and Choropleth Pattern Characteristics, The American Cartographer, 3(2):169-175.

Muller, Jean-Claude. 1979. Perception of Continuously Shaded Maps, Annals of the Association of American Geographers, 69(2):240-249.

Noronha, Valerian T. 1987. Choropleth Mapping in a Microcomputer Environment: A Critical Evaluation of Some Commercial Implementations, The American Cartographer, 14(2):139-154.

Paslawski, Jacek. 1984. In Search of a General Idea of Class Selection for Choropleth Maps, International Yearbook of Cartography, 24:159-171.

Peterson, Michael P. 1979. An Evaluation of Unclassed Crossed-Line Choropleth Mapping, The American Cartographer, 6(1):21-37.

Robinson, Arthur H., Sale, Randall D. & Morrison, Joel L. 1978. Elements of Cartography, Fourth Edition, John Wiley & Sons, New York, 448 pages.

Scripter, M.W. 1970. Nested-Means Map Classes for Statistical Maps, Annals of the Association of American Geographers, 60:385-393.

Semple, R. Keith & Green, Milford B. 1984. Classification in Human Geography, Spatial Statistics and Models, D. Reidel Publishing Co, Dordrecht, Netherlands: 55-79.

Silk, John, 1979. Statistical Concepts in Geography, George Allen & Unwin, London, G.B., 276 pages.

Smith, Richard M. 1986. Comparing Traditional Methods for Selecting Class Intervals on Choropleth Maps, Professional Geographer, 38(1):62-67.

Stefanovic, Pavao & Vries-Baayens, Annelieke. 1984. Classification Systems, Choropleth Maps and the Computer, ITC Journal, (1):52-57.

Stegena, Lajos & Csillag, Ferenc. 1987. Statistical Determination of Class Intervals for Maps, The Cartographic Journal, 24(2):142-146.

Tobler, W.R. 1973. Choropleth Maps Without Class Intervals?, Geographical Analysis, 5(3):262-265.

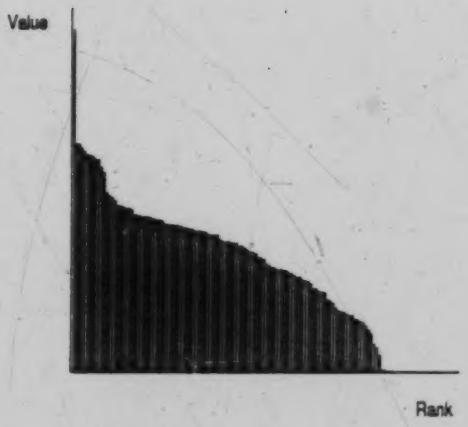
Truran, H.C. 1975. A Practical Guide to Statistical Maps and Diagrams, Heinemann Educational books, London, G.B., 60 pages.

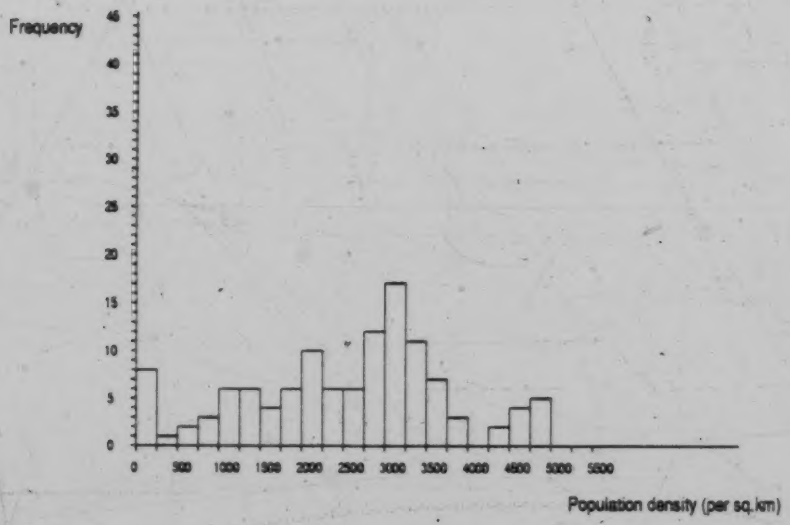
Turner, Eugene. 1987. MacChoro/Image Mapping Systems (software review), The American Cartographer, 14(1):69-71.

Appendix A

Rank-size graph, frequency histogram and table of results for each variable.

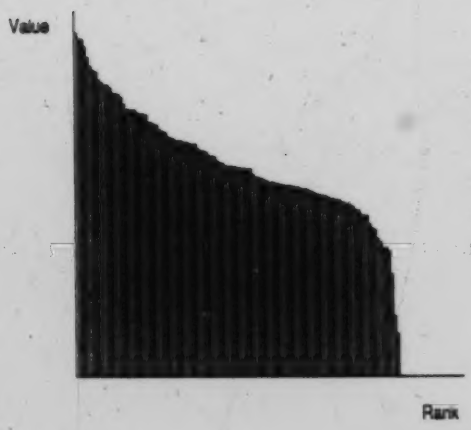
1. Population density

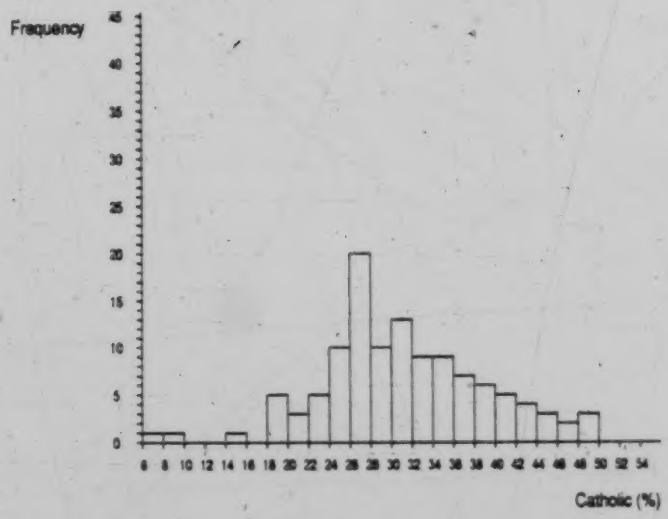




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.256	3.643	.689	.833	87.314	95.251
Equal Steps	.138	3.336	.682	.832	88.809	96.406
Arithmetic Progressions	7.126	6.475	.666	.829	88.023	96.568
Geometric Progressions	4.085	3.232	.603	.763	79.773	92.255
Reciprocal Progressions	1.817	1.974	.496	.634	65.808	77.168
Nested-Means	.208	2.944	.523	.669	69.200	80.177
Standard Deviations	.263	.399	.545	.674	71.985	81.906
"Optimal" (Variance)	3.728	19.993	.761	.891	92.465	98.929
"Optimal" (Abs.Dev.)	3.740	4.119	.761	.894	92.450	98.839

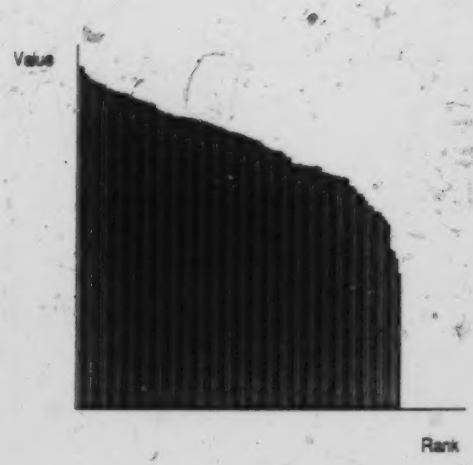
2. Catholic

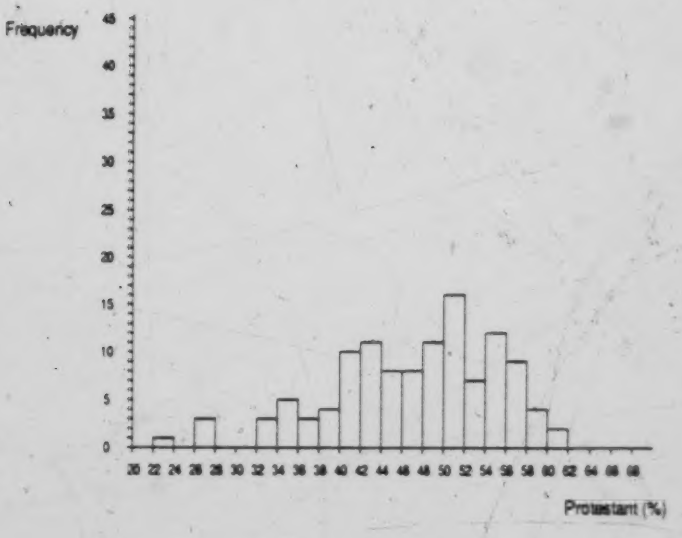




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.304	.170	.690	.815	86.516	94.355
Equal Steps	.088	.137	.678	.827	88.574	96.016
Arithmetic Progressions	.150	.510	.638	.802	86.589	95.559
Geometric Progressions	.483	.479	.617	.786	85.001	95.149
Reciprocal Progressions	.450	.604	.523	.727	74.257	90.431
Nested-Means	.316	.173	.543	.740	76.031	91.159
Standard Deviations	.222	.072	.559	.738	77.686	91.458
"Optimal" (Variance)	.236	.351	.683	.877	91.562	98.488
"Optimal" (Abs.Dev.)	.179	.349	.712	.877	90.913	98.486

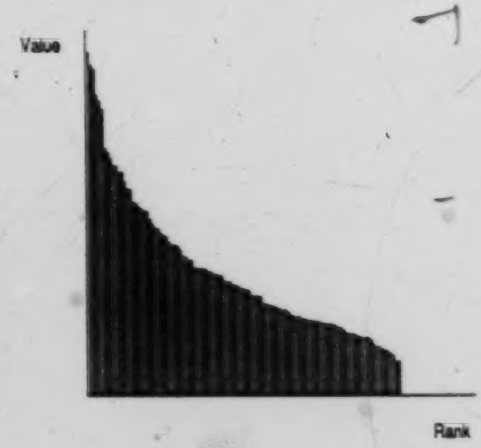
3. Protestant

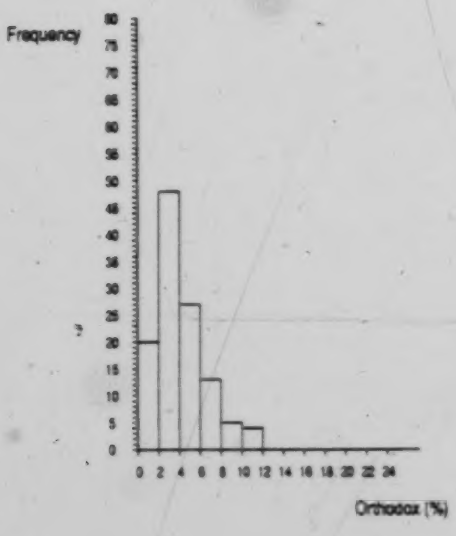




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.046	.023	.739	.836	90.079	96.268
Equal Steps	.017	.015	.746	.852	92.164	97.340
Arithmetic Progressions	.023	.066	.705	.833	90.360	96.885
Geometric Progressions	.018	.104	.691	.832	90.005	96.976
Reciprocal Progressions	.019	.087	.676	.824	89.018	96.720
Nested-Means	.036	.008	.682	.827	89.242	96.787
Standard Deviations	.039	.019	.685	.815	89.288	96.341
"Optimal" (Variance)	.026	.024	.756	.895	94.512	98.819
"Optimal" (Abs.Dev.)	.030	.024	.763	.895	92.939	98.819

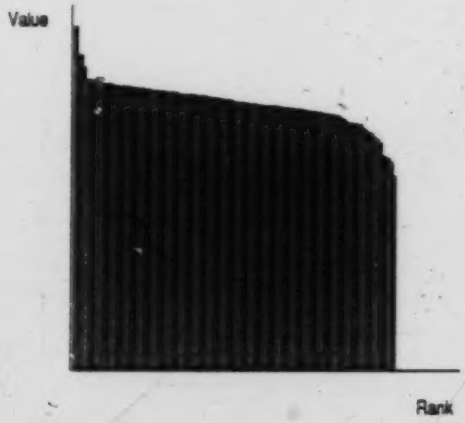
4. Orthodox

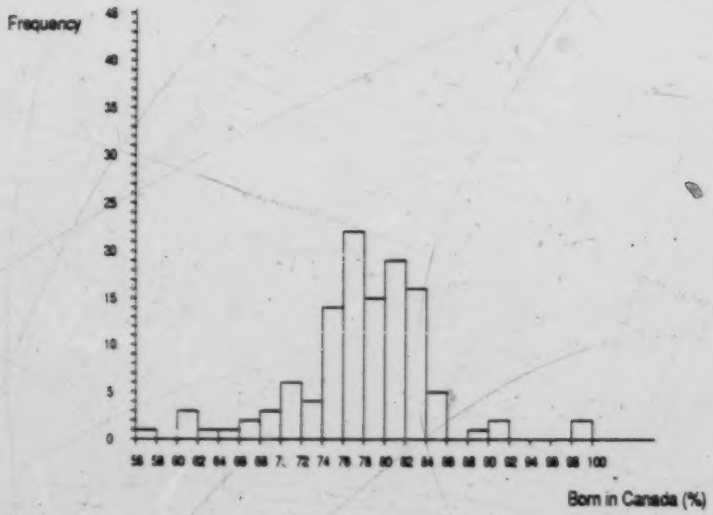




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.607	.173	.694	.821	86.863	95.238
Equal Steps	-	-	.696	.839	89.604	96.759
Arithmetic Progressions	-	-	.688	.841	89.783	97.123
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.655	.792	82.189	86.205
Standard Deviations	-	-	.660	.751	84.033	87.965
"Optimal" (Variance)	-	-	.725	.885	93.740	98.834
"Optimal" (Abs.Dev.)	-	-	.736	.887	93.024	98.678

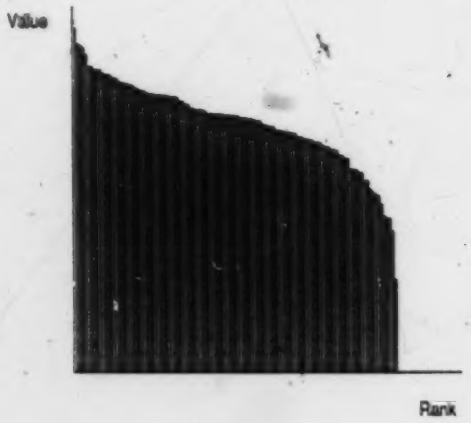
5. Born in Canada

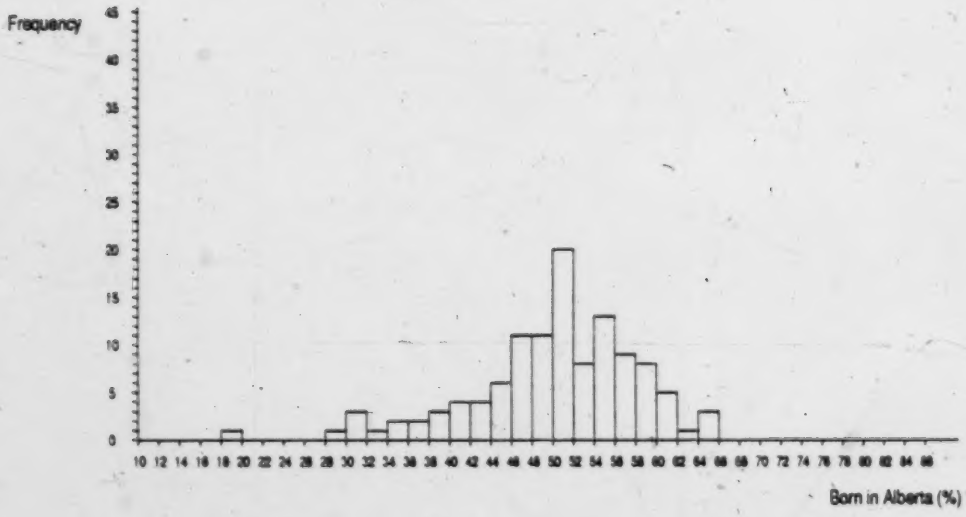




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.025	.010	.572	.744	75.686	87.774
Equal Steps	.006	.005	.571	.775	81.369	92.472
Arithmetic Progressions	.007	.035	.534	.762	80.831	93.281
Geometric Progressions	.009	.005	.559	.766	83.052	94.094
Reciprocal Progressions	.006	.008	.555	.768	83.522	94.568
Nested-Means	.024	.006	.554	.772	82.384	94.532
Standard Deviations	.009	.006	.555	.762	82.454	94.359
"Optimal" (Variance)	.009	.012	.670	.852	91.763	98.366
"Optimal" (Abs.Dev.)	.005	.009	.67[illegible]	.861	91.393	98.358

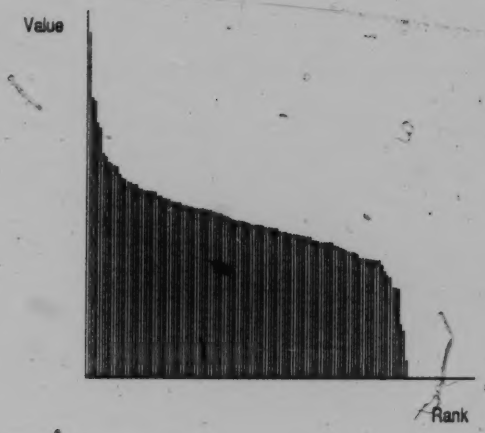
6. Born in Alberta

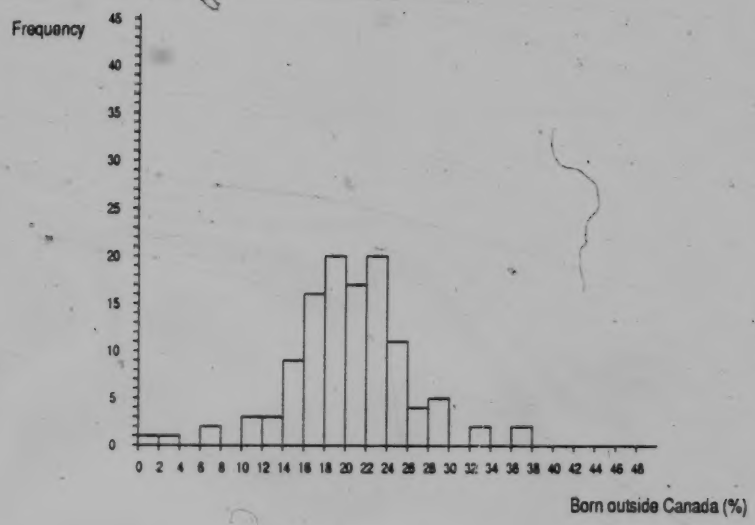




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.135	.088	.682	.814	85.635	94.255
Equal Steps	.025	.055	.649	.825	87.281	96.002
Arithmetic Progressions	.081	.365	.616	.784	85.602	94.879
Geometric Progressions	.098	.292	.613	.778	85.883	95.018
Reciprocal Progressions	.244	.292	.568	.754	82.046	93.915
Nested-Means	.112	.052	.581	.762	82.805	94.183
Standard Deviations	.117	.073	.587	.753	83.340	93.996
"Optimal" (Variance)	.050	.121	.680	.870	91.283	98.448
"Optimal" (Abs.Dev.)	.106	.121	.721	.870	89.862	98.435

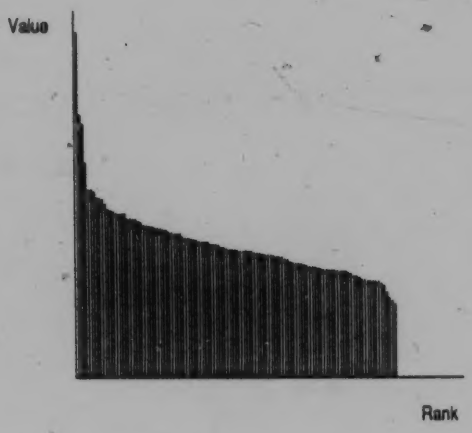
7. Born outside Canada

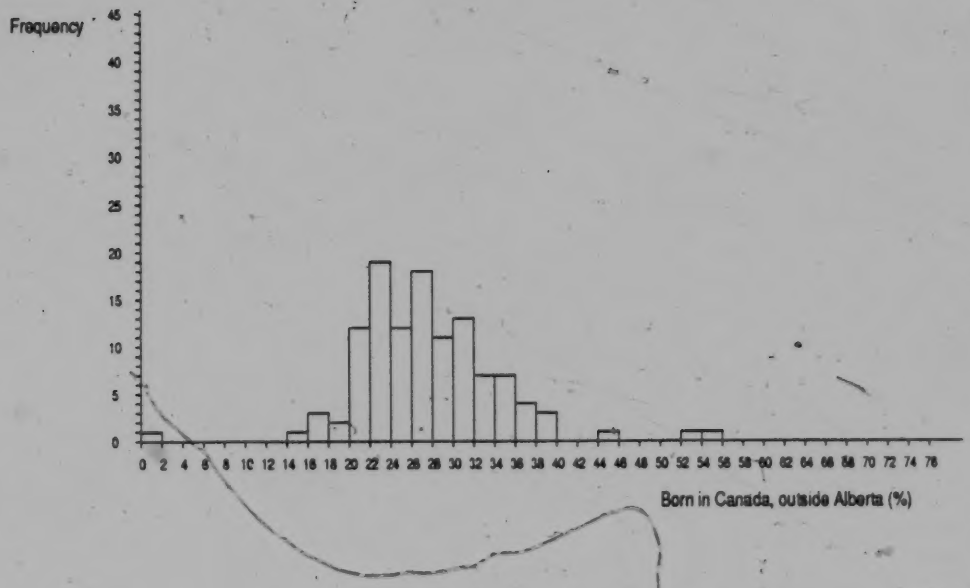




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.838	.474	.579	.744	74.325	86.678
Equal Steps	-	-	.561	.755	79.913	91.477
Arithmetic Progressions	-	-	.543	.736	80.115	92.093
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.784	.845	93.216	95.069
Standard Deviations	-	-	.813	.866	94.681	96.336
"Optimal" (Variance)	-	-	.641	.825	89.871	97.955
"Optimal" (Abs.Dev.)	-	-	.629	.844	83.152	97.816

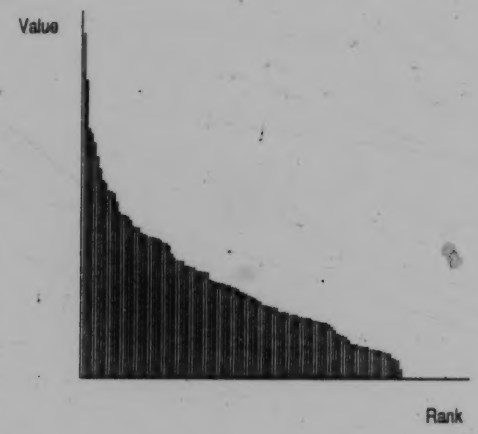
8. Born in Canada, outside Alberta

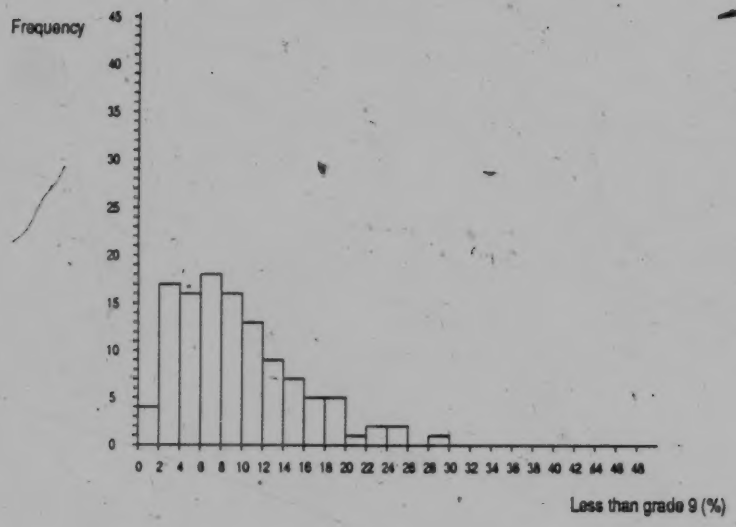




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	1.107	.945	.591	.716	67.548	78.607
Equal Steps	-	-	.557	.708	76.294	86.395
Arithmetic Progressions	-	-	.532	.689	77.896	88.378
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.790	.841	92.733	94.766
Standard Deviations	-	-	.817	.864	94.170	96.137
"Optimal" (Variance)	-	-	.628	.851	89.235	98.230
"Optimal" (Abs.Dev.)	-	-	.673	.854	84.765	98.109

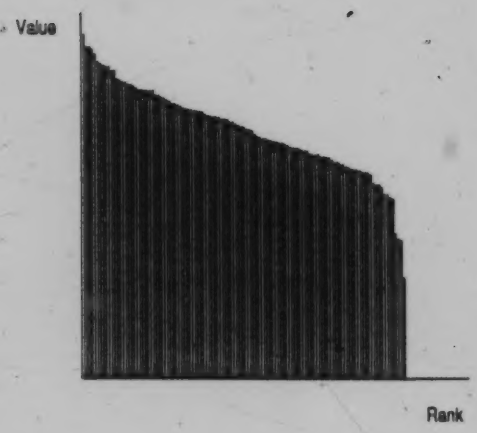
9. Less than grade 9

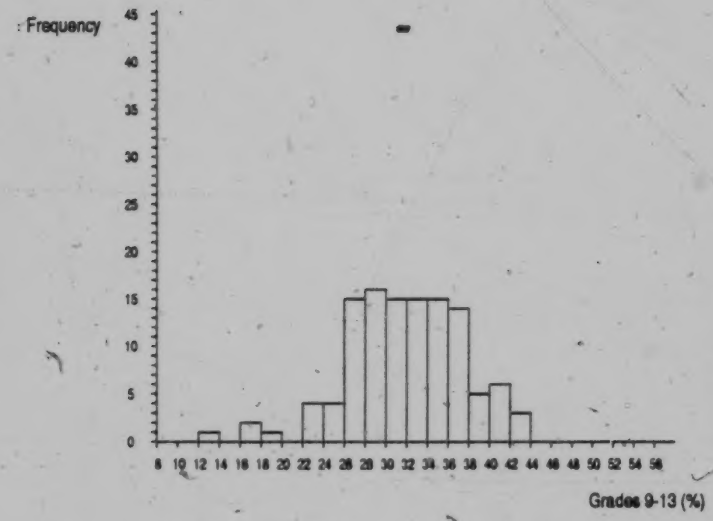




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.682	.740	.705	.838	84.690	93.436
Equal Steps	-	-	.688	.839	87.953	95.577
Arithmetic Progressions	-	-	.677	.839	88.517	96.253
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.709	.775	86.595	87.681
Standard Deviations	-	-	.722	.785	88.219	89.821
"Optimal" (Variance)	-	-	.712	.878	92.884	98.623
"Optimal" (Abs.Dev.)	-	-	.758	.884	90.822	98.527

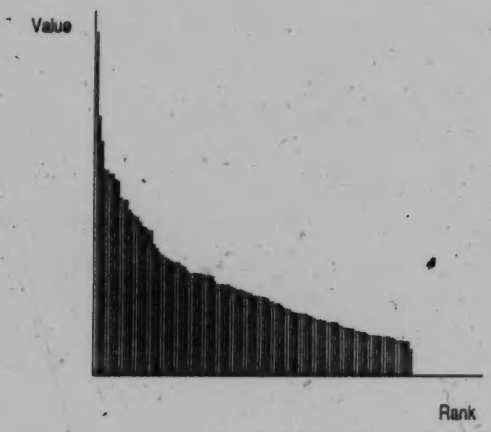
10. Grades 9-13

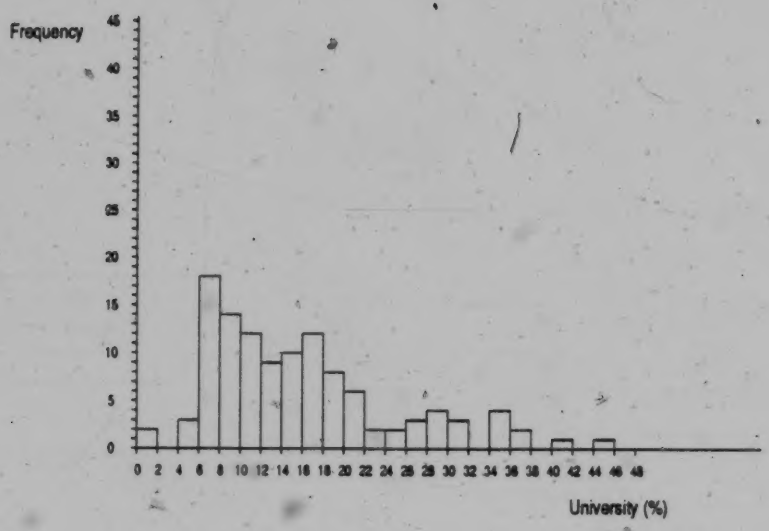




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.122	.075	.686	.827	86.873	94.488
Equal Steps	.042	.186	.689	.834	89.141	96.161
Arithmetic Progressions	.060	.349	.653	.807	87.008	95.651
Geometric Progressions	.066	.268	.638	.803	86.447	95.783
Reciprocal Progressions	.076	.299	.605	.781	83.378	94.348
Nested-Means	.122	.053	.614	.787	83.755	94.908
Standard Deviations	.085	.027	.618	.777	84.228	94.596
"Optimal" (Variance)	.049	.098	.717	.870	92.734	98.505
"Optimal" (Abs.Dev.)	.051	.047	.721	.878	92.372	98.369

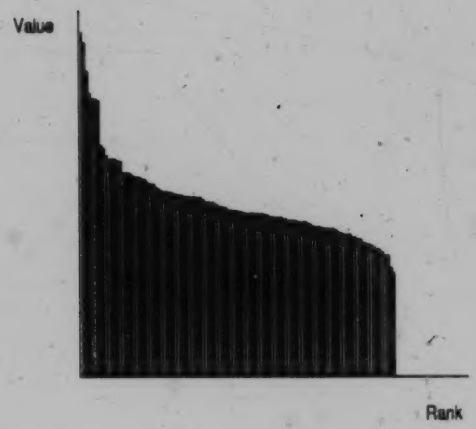
11. University

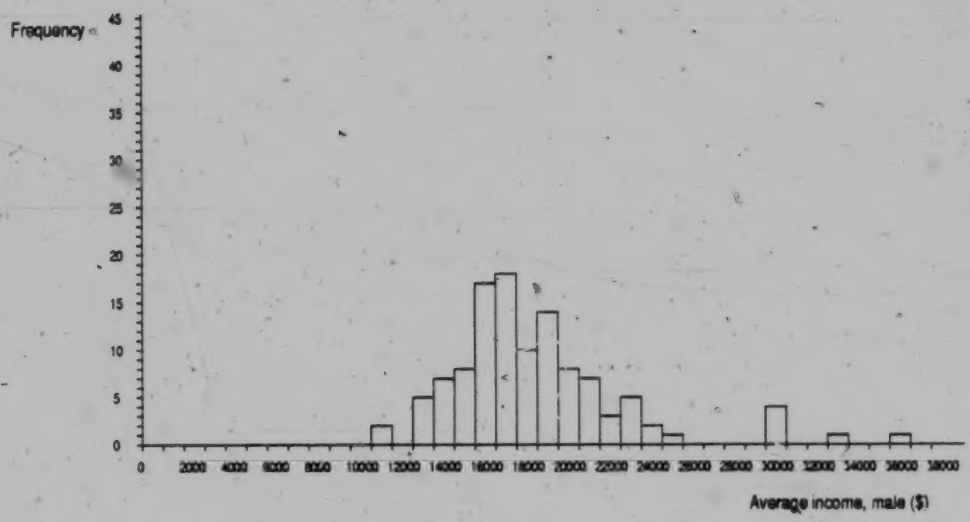




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.953	.767	.715	.823	84.165	92.187
Equal Steps	-	-	.699	.810	87.857	94.512
Arithmetic Progressions	-	-	.693	.816	88.918	95.549
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.728	.782	86.593	88.604
Standard Deviations	-	-	.741	.791	88.346	90.577
"Optimal" (Variance)	-	-	.711	.878	92.805	98.690
"Optimal" (Abs.Dev.)	-	-	.741	.877	91.000	98.049

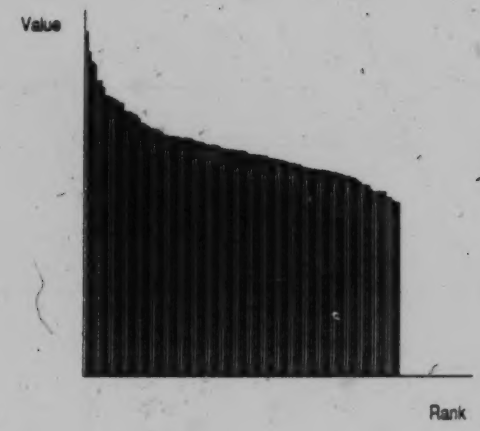
12. Average income, male

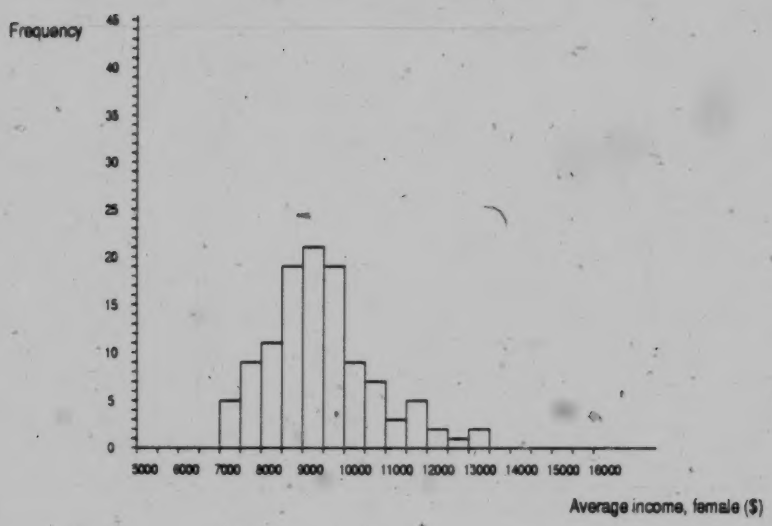




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.924	.890	.597	.818	74.641	93.977
Equal Steps	2.044	2.546	.552	.783	78.417	94.790
Arithmetic Progressions	2.836	3.670	.502	.731	76.546	93.227
Geometric Progressions	3.440	3.690	.387	.629	59.034	85.995
Reciprocal Progressions	1.573	2.283	.310	.503	47.227	68.796
Nested-Means	.900	.879	.358	.557	52.868	73.357
Standard Deviations	.923	3.356	.390	.580	57.031	76.345
"Optimal" (Variance)	-	-	-	-	-	-
"Optimal" (Abs.Dev.)	-	-	-	-	-	-

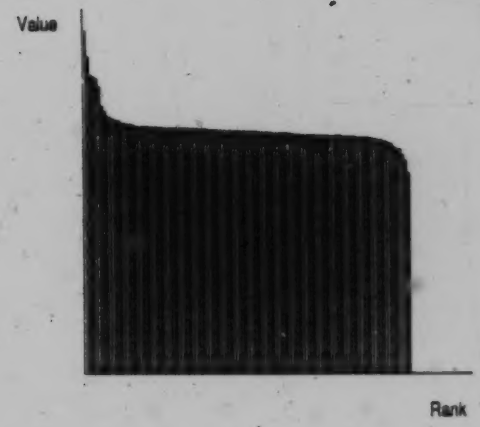
13. Average income, female

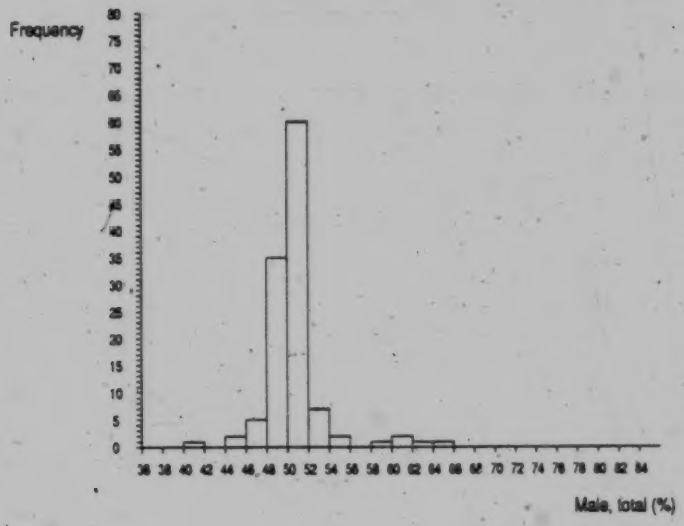




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.882	.882	.701	.835	86.558	96.069
Equal Steps	2.447	2.849	.593	.750	81.625	93.615
Arithmetic Progressions	3.039	3.267	.511	.673	74.422	89.172
Geometric Progressions	3.288	3.459	.383	.589	55.817	81.013
Reciprocal Progressions	1.865	2.391	.307	.471	44.653	64.811
Nested-Means	.883	.880	.364	.529	51.670	70.052
Standard Deviations	.889	2.337	.404	.556	56.747	73.367
"Optimal" (Variance)	-	-	-	-	-	-
"Optimal" (Abs.Dev.)	-	-	-	-	-	-

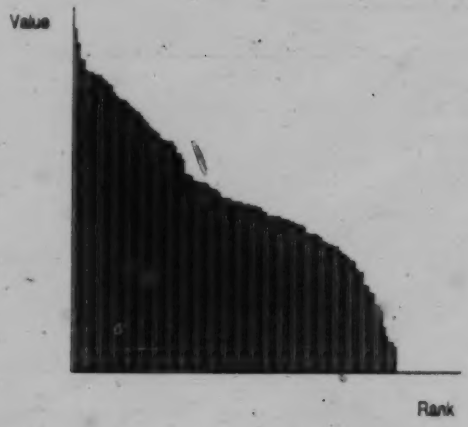
14. Male, total

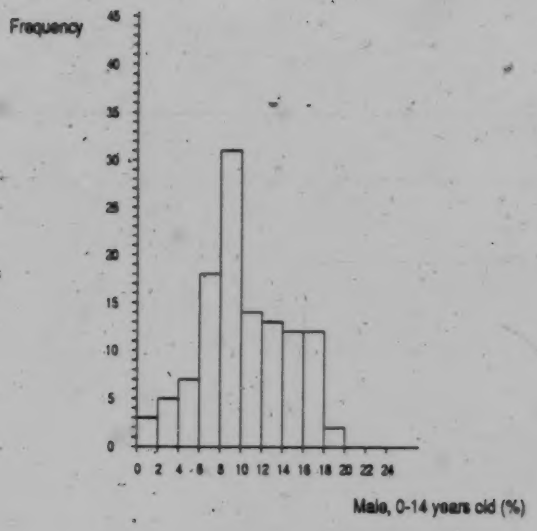




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.045	.022	.336	.595	47.696	70.276
Equal Steps	.008	.103	.422	.663	68.524	83.343
Arithmetic Progressions	.006	.034	.432	.679	73.883	87.544
Geometric Progressions	.009	.006	.455	.679	77.865	89.686
Reciprocal Progressions	.012	.009	.485	.697	80.520	91.277
Nested-Means	.018	.009	.490	.704	79.672	91.275
Standard Deviations	.010	.009	.499	.699	79.941	91.806
"Optimal" (Variance)	.017	.011	.621	.830	92.182	98.462
"Optimal" (Abs.Dev.)	.020	.014	.641	.836	88.505	98.367

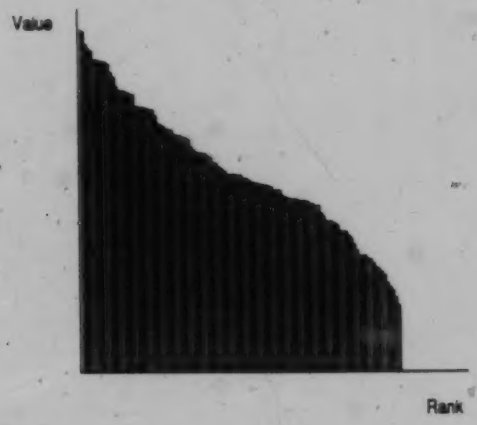
15. Male, 0-14 years old

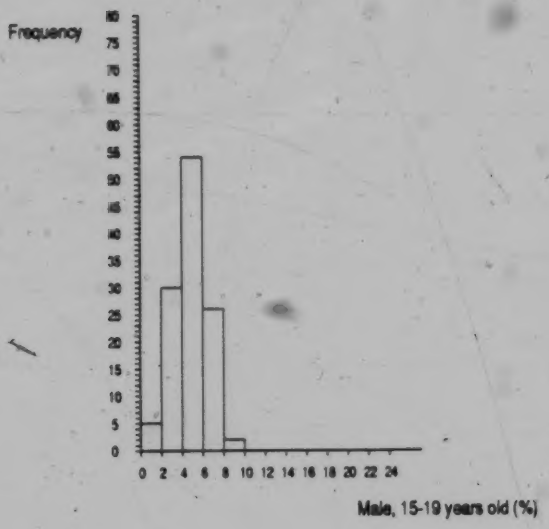




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.145	.053	.729	.853	91.723	97.521
Equal Steps	.083	.216	.745	.866	93.181	98.071
Arithmetic Progressions	.184	.074	.730	.852	92.399	97.715
Geometric Progressions	.132	.082	.716	.846	91.535	97.497
Reciprocal Progressions	.183	.442	.608	.762	80.215	90.941
Nested-Means	.210	.040	.620	.777	81.796	92.031
Standard Deviations	.049	.109	.629	.769	83.107	92.173
"Optimal" (Variance)	1.762	.569	.614	.757	47.374	67.020
"Optimal" (Abs.Dev.)	1.729	1.707	.613	.710	47.637	50.609

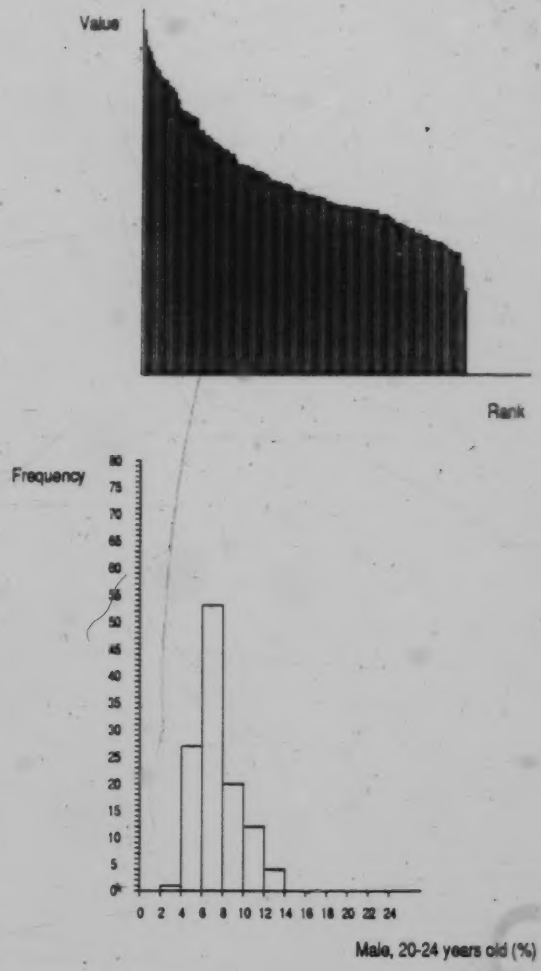
16. Male, 15-19 years old





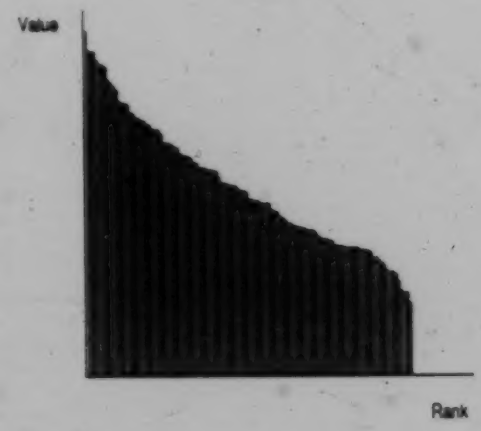
Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.626	.659	.742	.851	91.847	96.913
Equal St[illegible]s	-	-	.725	.847	92.077	97.378
Arithmetic Progressions	-	-	.690	.825	89.987	96.805
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.733	.801	88.692	91.036
Standard Deviations	-	-	.758	.815	90.878	92.797
"Optimal" (Variance)	-	-	.330	.423	18.917	19.839
"Optimal" (Abs.Dev.)	-	-	.332	.424	18.902	19.788

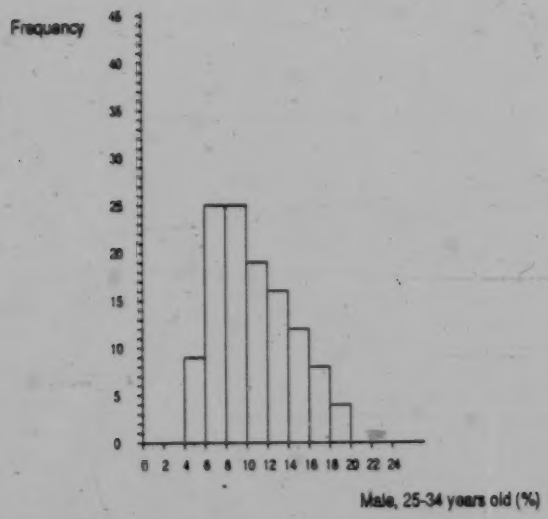
17. Male, 20-24 years old



Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of Classes	5	10	5	10	5	10
Quantiles	.083	.058	.734	.863	91.001	97.315
Equal Steps	.069	.018	.732	.859	91.967	97.761
Arithmetic Progressions	.069	.124	.717	.848	91.364	97.562
Geometric Progressions	.027	.027	.715	.846	91.391	97.532
Reciprocal Progressions	.044	.200	.693	.836	89.646	97.016
Nested-Means	.079	.061	.695	.839	89.900	97.071
Standard Deviations	.077	.067	.695	.824	89.988	96.592
"Optimal" (Variance)	2.326	1.060	.456	.576	27.044	38.905
"Optimal" (Abs.Dev.)	2.427	1.875	.459	.578	26.385	38.879

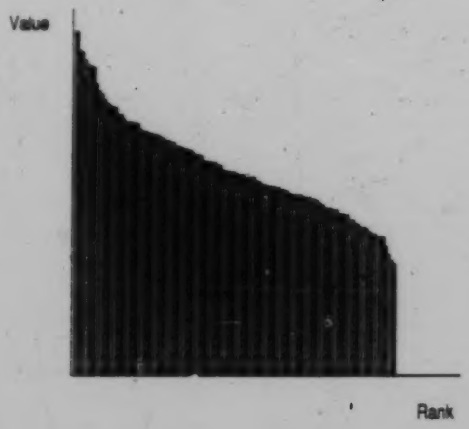
18. Male, 25-34 years old

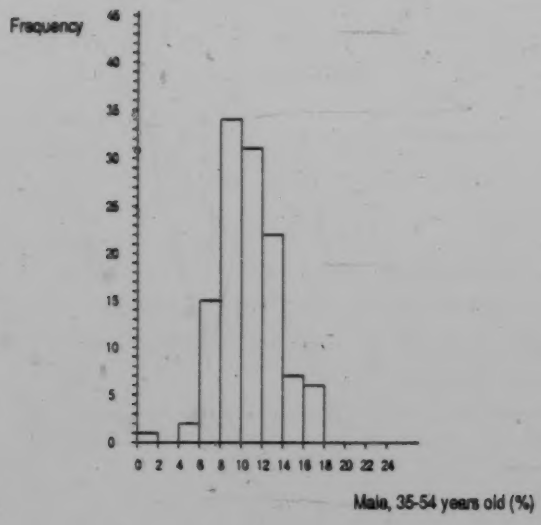




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.071	.024	.744	.880	92.852	98.234
Equal Steps	.058	.020	.749	.873	93.461	98.332
Arithmetic Progressions	.046	.015	.734	.864	92.618	98.080
Geometric Progressions	.037	.015	.733	.866	92.504	98.092
Reciprocal Progressions	.060	.025	.713	.854	90.446	97.434
Nested-Means	.081	.021	.711	.856	90.499	97.497
Standard Deviations	.050	.044	.706	.834	90.414	96.649
"Optimal" (Variance)	.044	.030	.765	.892	94.745	98.806
"Optimal" (Abs.Dev.)	.074	.021	.774	.895	94.238	98.673

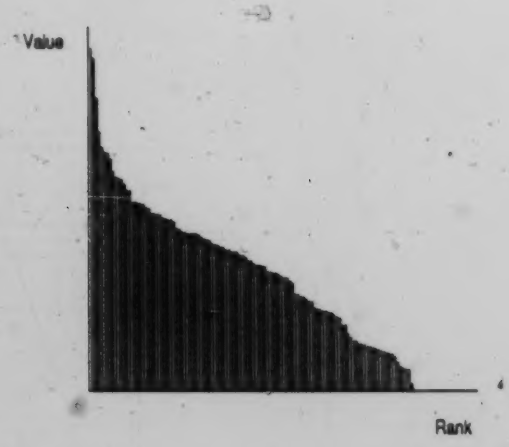
19. Male, 35-54 years old

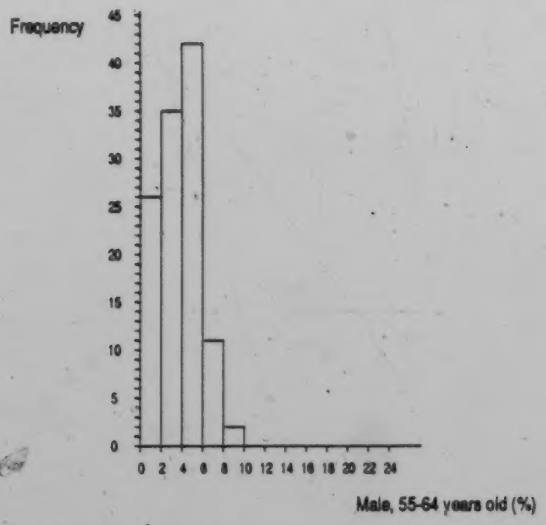




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.821	.784	.694	.821	85.330	92.720
Equal Steps	-	-	.666	.804	86.897	94.595
Arithmetic Progressions	-	-	.605	.774	83.057	94.027
Geometric Progressions	-	-		-		-
Reciprocal Progressions	-	-		-		-
Nested-Means	-	-	.792	.848	93.898	95.560
Standard Deviations	-	-	.822	.868	95.306	96.675
"Optimal" (Variance)	-	-	.701	.864	92.038	98.362
"Optimal" (Abs.Dev.)	-	-	.923	.867	88.729	98.109

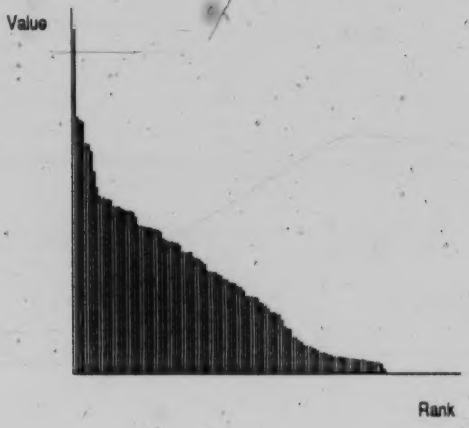
20. Male, 55-64 years old

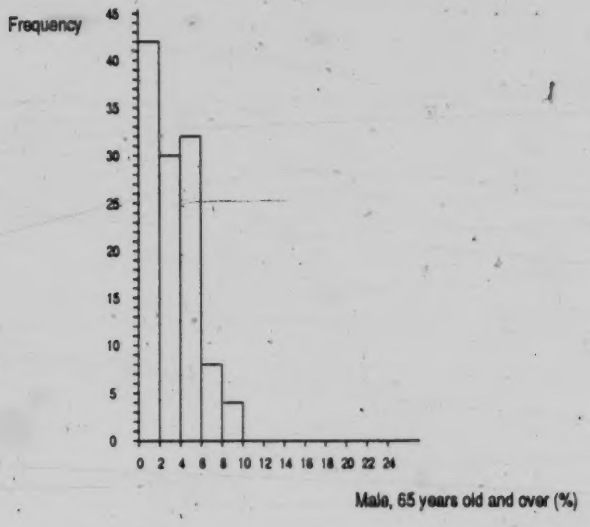




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.323	.325	.730	.839	88.591	94.898
Equal Steps	.125	.344	.720	.845	90.550	96.441
Arithmetic Progressions	.291	.470	.706	.845	90.256	96.848
Geometric Progressions	.771	.658	.675	.826	87.333	96.049
Reciprocal Progressions	1.409	1.364	.561	.734	72.601	86.090
Nested-Means	.153	.380	.581	.749	74.998	87.691
Standard Deviations	.278	.226	.595	.746	76.943	88.524
"Optimal" (Variance)	.193	.264	.701	.846	87.840	91.514
"Optimal" (Abs.Dev.)	.570	.264	.701	.846	79.299	91.514

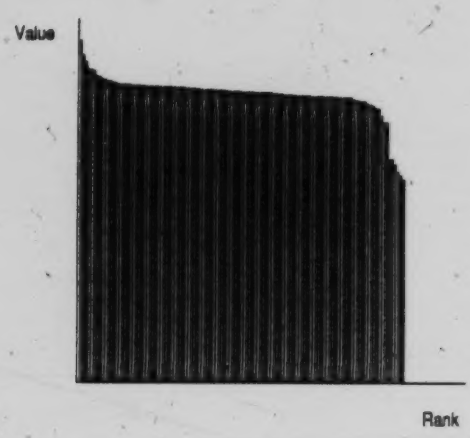
21. Male, 65 years old and over

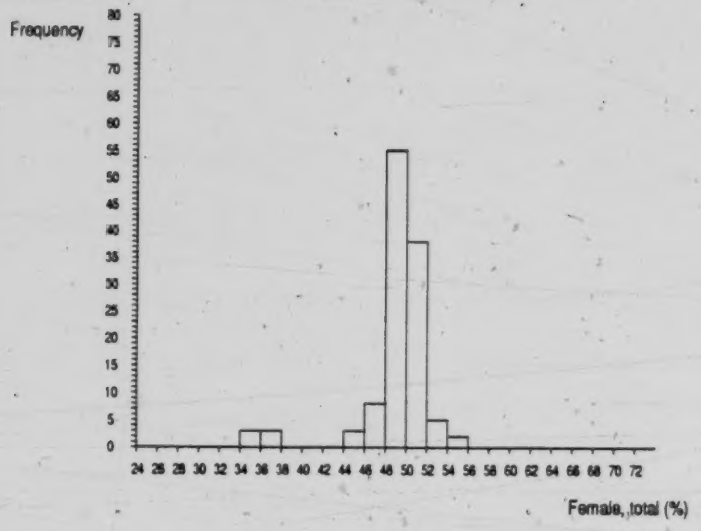




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.781	.225	.740	.867	89.181	95.052
Equal Steps	-	-	.742	.867	91.305	96.646
Arithmetic Progressions	-	-	.745	.868	91.641	97.154
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.742	.782	87.601	86.271
Standard Deviations	-	-	.737	.779	87.928	87.551
"Optimal" (Variance)	-	-	.764	.884	88.453	97.398
"Optimal" (Abs.Dev.)	-	-	.765	.884	88.371	97.246

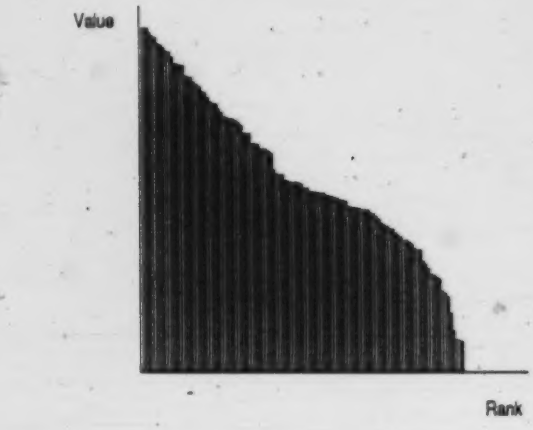
22. Female, total

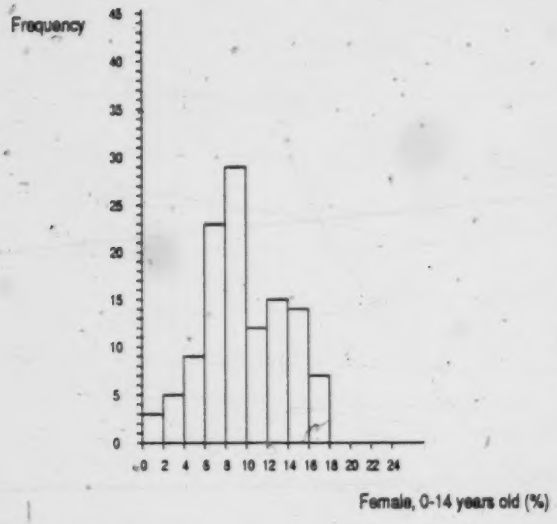




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.029	.006	.369	.645	53.559	79.367
Equal Steps	.123	.120	.463	.680	71.705	87.983
Arithmetic Progressions	.129	.157	.482	.682	75.986	90.303
Geometric Progressions	.110	.110	.495	.702	79.523	92.002
Reciprocal Progressions	.014	.102	.507	.696	81.513	92.680
Nested-Means	.018	.011	.512	.714	81.553	93.342
Standard Deviations	.010	.092	.518	.712	81.959	93.676
"Optimal" (Variance)	.023	.018	.703	.862	94.913	98.937
"Optimal" (Abs.Dev.)	.023	.018	.704	.861	94.553	98.551

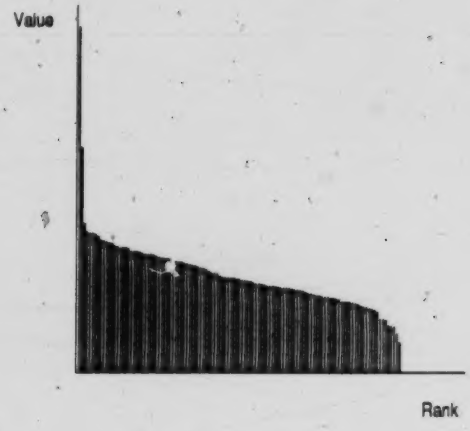
23. Female, 0-14 years old

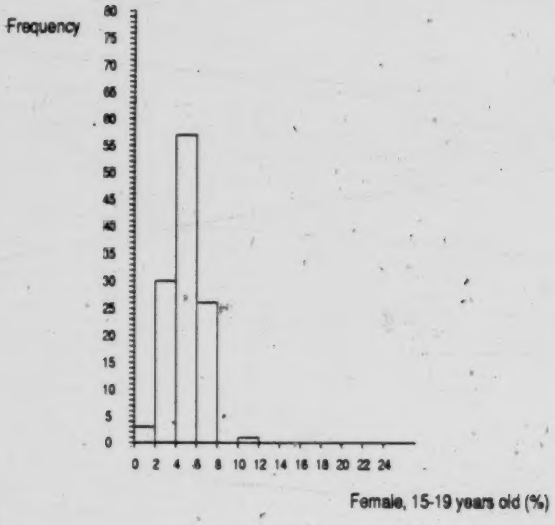




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.162	.036	.759	.863	93.176	97.828
Equal Steps	.178	.144	.765	.869	94.052	98.195
Arithmetic Progressions	.138	.045	.738	.859	92.590	97.919
Geometric Progressions	.141	.083	.716	.850	91.047	97.577
Reciprocal Progressions	.095	.474	.617	.773	80.595	91.377
Nested-Means	.178	.056	.634	.787	82.389	92.410
Standard Deviations	.022	.127	.645	.780	83.769	92.547
"Optimal" (Variance)	1.517	1.205	.652	.740	54.224	60.881
"Optimal" (Abs.Dev.)	1.365	1.343	.652	.741	54.386	59.245

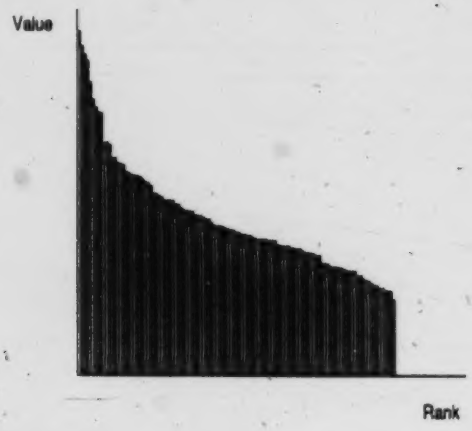
24. Female, 15-19 years old

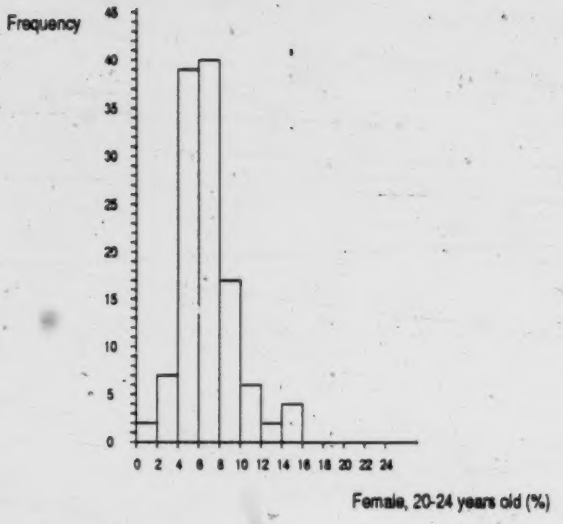




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	1.331	1.024	.604	.719	62.004	69.944
Equal Steps	-	-	.496	.708	69.263	82.053
Arithmetic Progressions	-	-	.453	.705	71.135	86.098
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.652	.764	80.210	86.792
Standard Deviations	-	-	.684	.786	82.729	89.523
"Optimal" (Variance)	-	-	.648	.862	90.989	98.644
"Optimal" (Abs.Dev.)	-	-	.676	.860	87.470	97.794

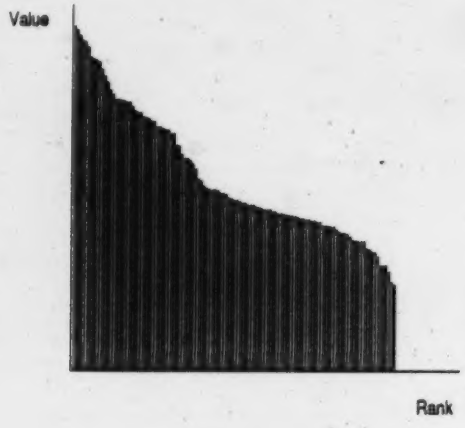
25. Female, 20-24 years old

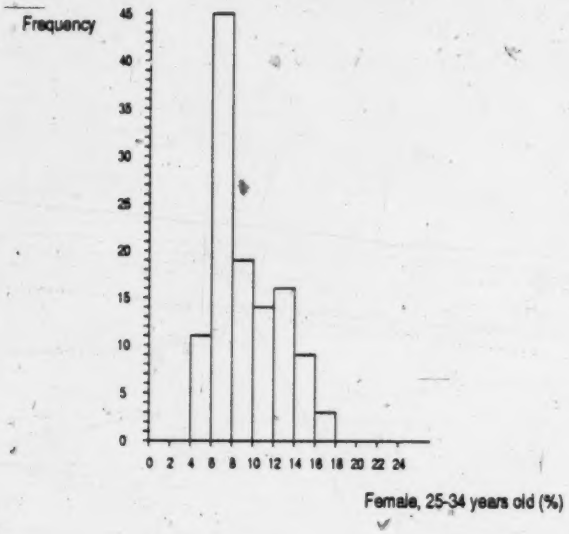




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.821	.678	.640	.791	81.490	92.469
Equal Steps	-	-	.635	.799	85.569	94.855
Arithmetic Progressions	-	-	.610	.787	85.252	95.106
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.726	.801	88.387	91.494
Standard Deviations	-	-	.750	.818	90.459	93.248
"Optimal" (Variance)	-	-	.700	.869	91.432	98.046
"Optimal" (Abs.Dev.)	-	-	.722	.870	89.924	98.043

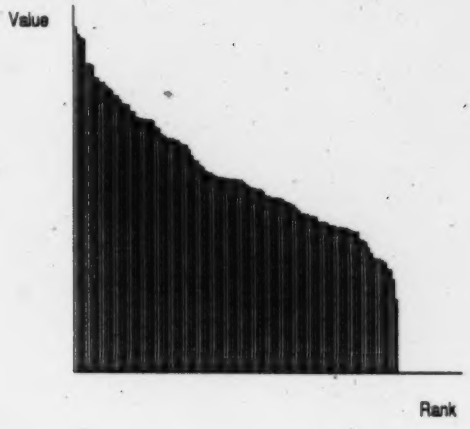
26. Female, 25-34 years old

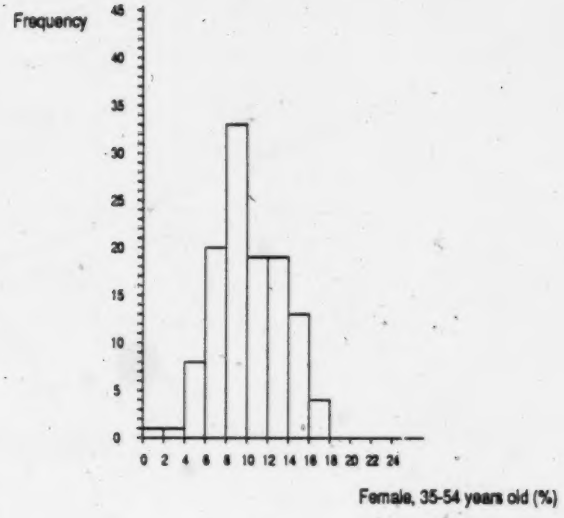




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.042	.010	.758	.891	92.732	98.343
Equal Steps	.041	.018	.774	.892	94.089	98.594
Arithmetic Progressions	.035	.019	.762	.886	93.563	98.469
Geometric Progressions	.035	.025	.762	.887	93.544	98.516
Reciprocal Progressions	.044	.018	.750	.876	92.342	97.953
Nested-Means	.066	.014	.748	.876	92.329	98.007
Standard Deviations	.042	.042	.744	.860	92.259	97.457
"Optimal" (Variance)	.040	.023	.795	.904	95.403	98.871
"Optimal" (Abs.Dev.)	.050	.025	.797	.905	95.287	98.854

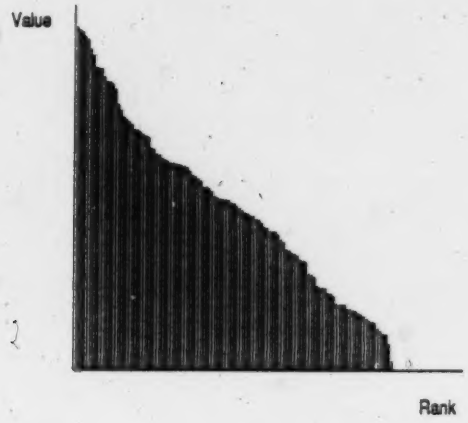
27. Female, 35-54 years old

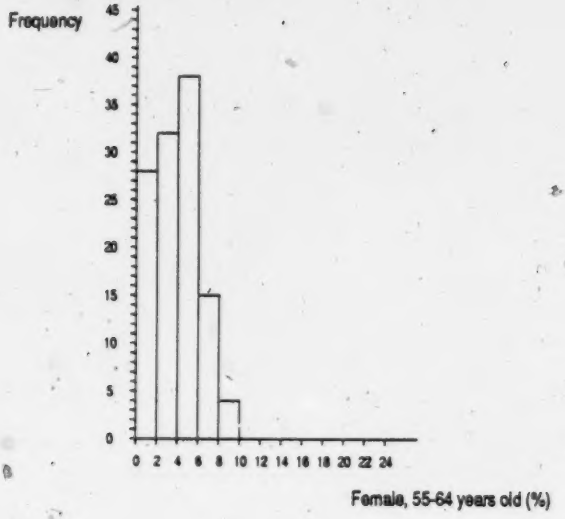




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.781	.716	.724	.833	89.418	95.334
Equal Steps	-	-	.698	.837	89.891	96.534
Arithmetic Progressions	-	-	.658	.817	87.410	96.013
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.776	.836	92.624	94.376
Standard Deviations	-	-	.803	.854	94.253	95.695
"Optimal" (Variance)	-	-	.747	.877	92.536	98.520
"Optimal" (Abs.Dev.)	-	-	.749	.883	92.115	98.407

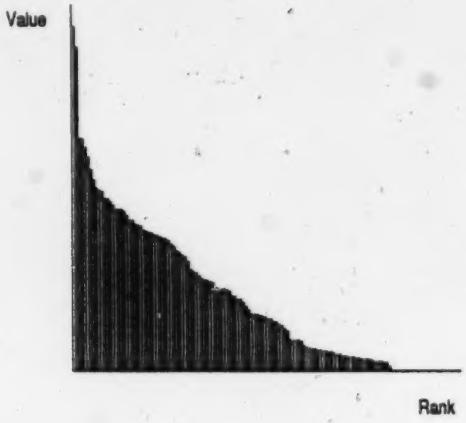
28. Female, 55-64 years old

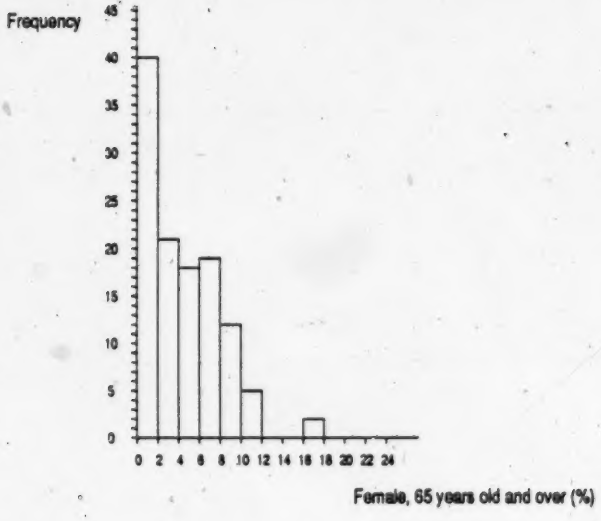




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.402	.329	.759	.864	93.585	98.087
Equal Steps	-	-	.759	.871	94.103	98.347
Arithmetic Progressions	-	-	.740	.867	93.122	98.193
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.705	.777	85.273	87.234
Standard Deviations	-	-	.718	.779	87.242	88.896
"Optimal" (Variance)	-	-	.723	.829	81.919	86.974
"Optimal" (Abs.Dev.)	-	-	.733	.830	81.310	86.953

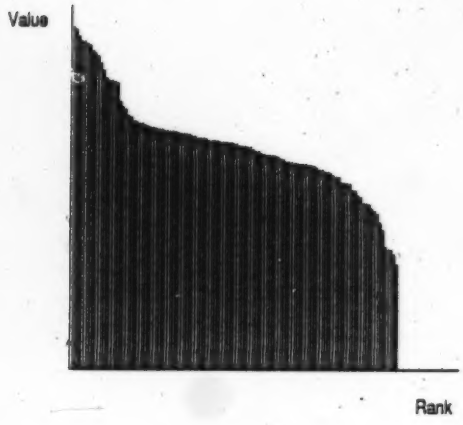
29. Female, 65 years old and over

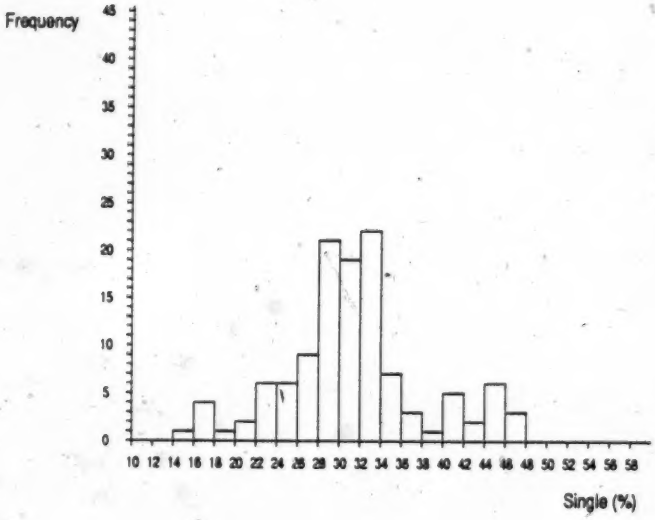




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.668	.661	.736	.853	87.042	93.460
Equal Steps	-	-	.723	.864	89.787	95.487
Arithmetic Progressions	-	-	.729	.869	90.571	96.763
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.727	.784	86.625	86.892
Standard Deviations	-	-	.720	.779	87.084	88.263
"Optimal" (Variance)	-	-	.792	.909	95.786	99.186
"Optimal" (Abs.Dev.)	-	-	.777	.910	92.128	99.179

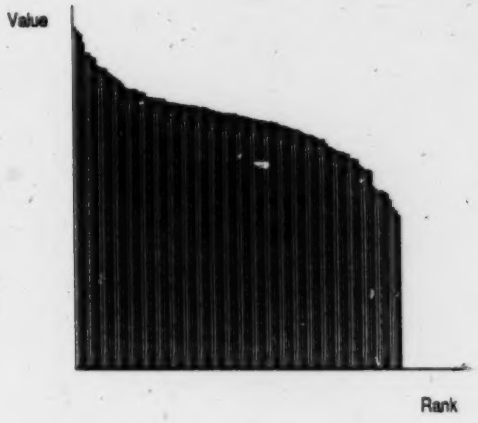
30. Single

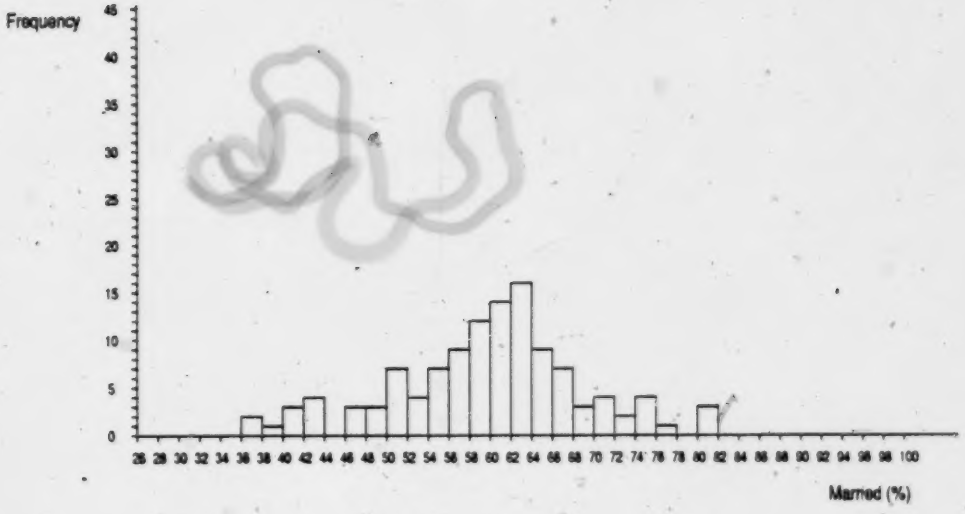




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.040	.008	.649	.811	85.269	96.049
Equal Steps	.023	.009	.655	.826	88.683	97.134
Arithmetic Progressions	.023	.019	.646	.819	88.994	97.070
Geometric Progressions	.035	.016	.661	.823	90.020	97.270
Reciprocal Progressions	.035	.020	.623	.811	87.315	96.871
Nested-Means	.042	.013	.628	.813	87.522	96.926
Standard Deviations	.034	.031	.634	.804	88.042	96.648
"Optimal" (Variance)	.041	.025	.733	.871	94.148	98.716
"Optimal" (Abs.Dev.)	.038	.024	.735	.880	94.124	98.663

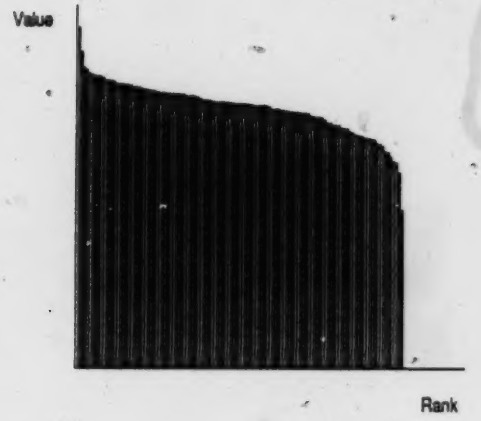
31. Married

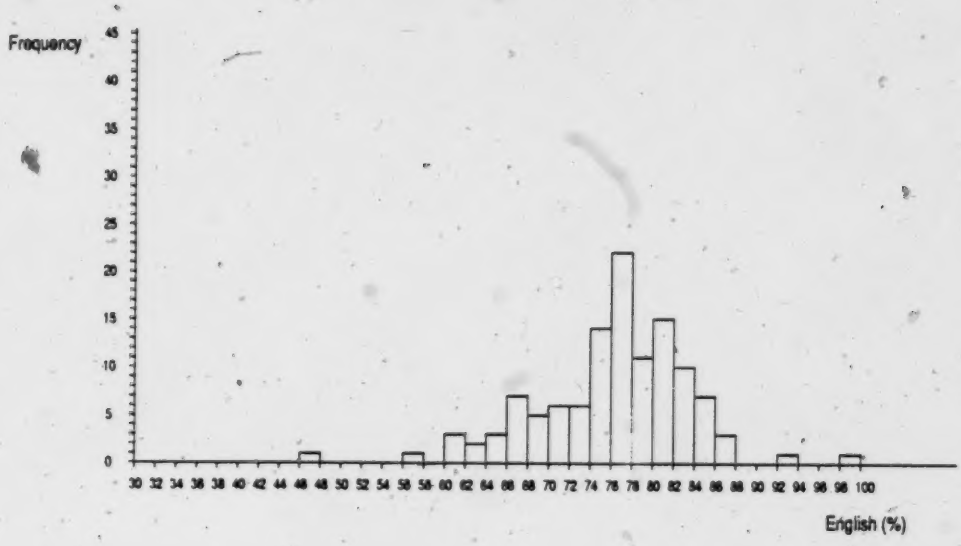




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.019	.008	.665	.836	87.785	96.652
Equal Steps	.006	.007	.675	.841	90.064	97.394
Arithmetic Progressions	.008	.007	.660	.828	89.294	97.130
Geometric Progressions	.015	.008	.664	.828	89.769	97.265
Reciprocal Progressions	.016	.072	.658	.822	89.387	97.150
Nested-Means	.020	.007	.653	.821	88.924	97.086
Standard Deviations	.013	.009	.650	.802	88.911	96.414
"Optimal" (Variance)	.013	.013	.717	.864	93.375	98.510
"Optimal" (Abs.Dev.)	.012	.012	.723	.867	93.181	98.270

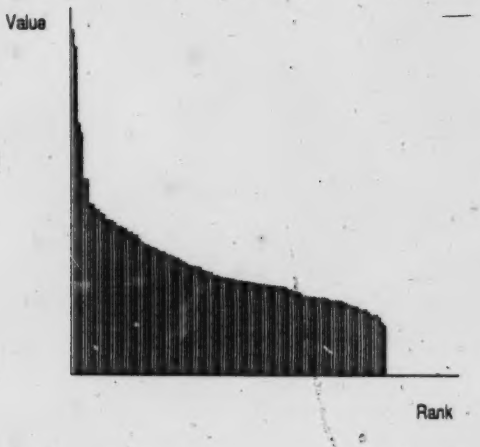
32. English

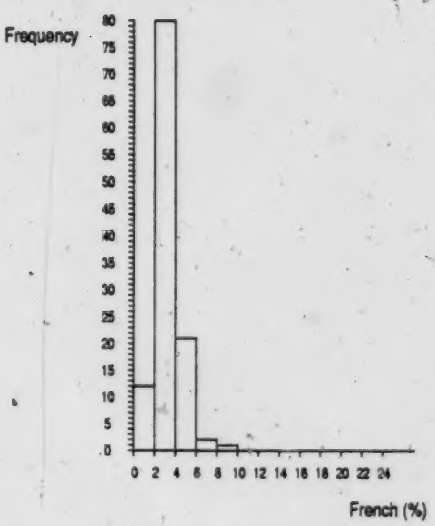




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.056	.038	.680	.797	83.028	90.481
Equal Steps	.037	.110	.632	.789	84.601	93.401
Arithmetic Progressions	.159	.176	.584	.754	81.820	93.052
Geometric Progressions	.026	.086	.569	.755	81.896	93.622
Reciprocal Progresssions	.032	.127	.561	.748	81.389	93.638
Nested-Means	.055	.028	.571	.757	81.500	93.415
Standard Deviations	.050	.155	.573	.748	81.599	93.333
"Optimal" (Variance)	.025	.028	.657	.863	89.565	98.077
"Optimal" (Abs.Dev.)	.054	.036	.687	.862	83.688	97.883

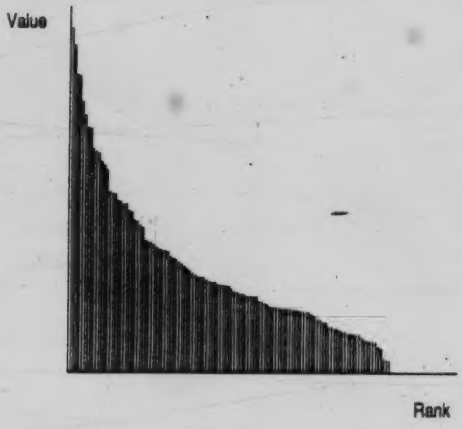
33. French

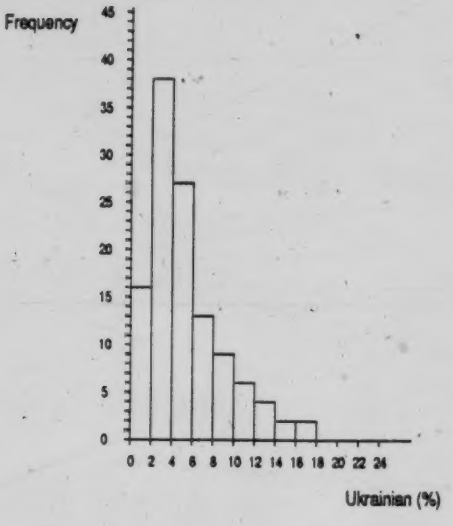




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	1.004	.499	.579	.720	69.434	81.730
Equal Steps	-	-	.581	.755	78.307	89.393
Arithmetic Progressions	-	-	.589	.758	81.400	91.719
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.609	.715	76.760	83.752
Standard Deviations	-	-	.615	.724	78.456	86.067
"Optimal" (Variance)	-	-	.301	.475	30.166	43.703
"Optimal" (Abs.Dev.)	-	-	.310	.478	30.106	43.700

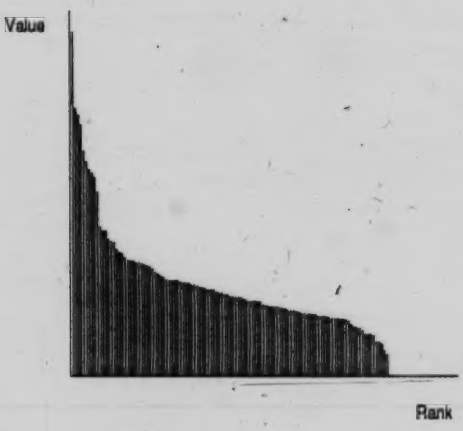
34. Ukrainian

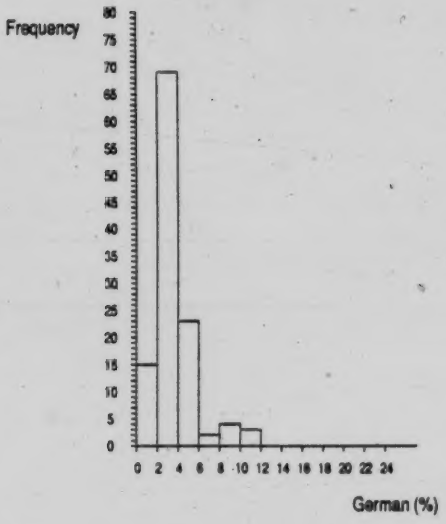




Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.708	.502	.714	.839	86.584	95.513
Equal Steps	-	-	.716	.841	90.065	96.819
Arithmetic Progressions	-	-	.711	.837	90.775	97.037
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.681	.752	83.796	86.200
Standard Deviations	-	-	.685	.753	85.511	88.085
"Optimal" (Variance)	-	-	.507	.669	36.683	55.743
"Optimal" (Abs.Dev.)	-	-	.514	.627	33.275	45.643

35. German





Method	Sum of Differences		Tabular Accuracy Index		Goodness of Variance Fit	
Number of classes	5	10	5	10	5	10
Quantiles	.910	.484	.566	.789	71.982	90.523
Equal Steps	-	-	.579	.789	80.291	93.794
Arithmetic Progressions	-	-	.588	.792	83.046	94.975
Geometric Progressions	-	-	-	-	-	-
Reciprocal Progressions	-	-	-	-	-	-
Nested-Means	-	-	.615	.726	78.538	84.060
Standard Deviations	-	-	.628	.729	80.853	86.178
"Optimal" (Variance)	-	-	.382	.648	34.879	61.792
"Optimal" (Abs.Dev.)	-	-	.356	.649	22.898	61.790